

How Do Genomes Differ?

Lecture : 13th October 2015

Dr CA Sargent

cas1001@cam.ac.uk

Aims of these lectures

- Why do genomic differences matter?
- Reminder of types and levels of variation
- How is this related to the products of the mammalian genome?
- How can variation be used in genetic studies?
- *Act as introduction and link to other lectures in this series*

Differences between species



Speke's: *Gazella spekei*

Some species are anatomically similar. Karyotypes allow a cytogenetic perspective that can distinguish on chromosome number or banding patterns.



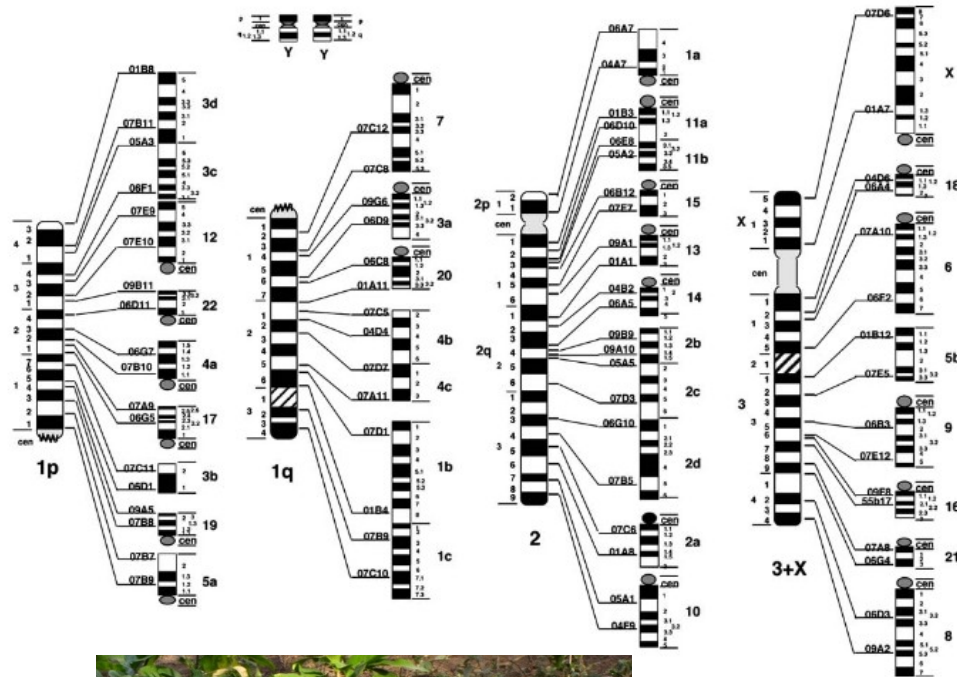
Mountain: *Gazella gazella*



Dorcas: *Gazella leptoceros*

Differences in Chromosome number or structure

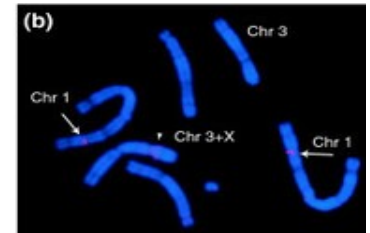
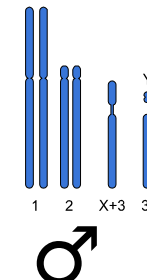
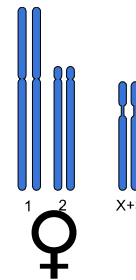
Adapted from Chi et al, Chromosoma (2005) 114: 167



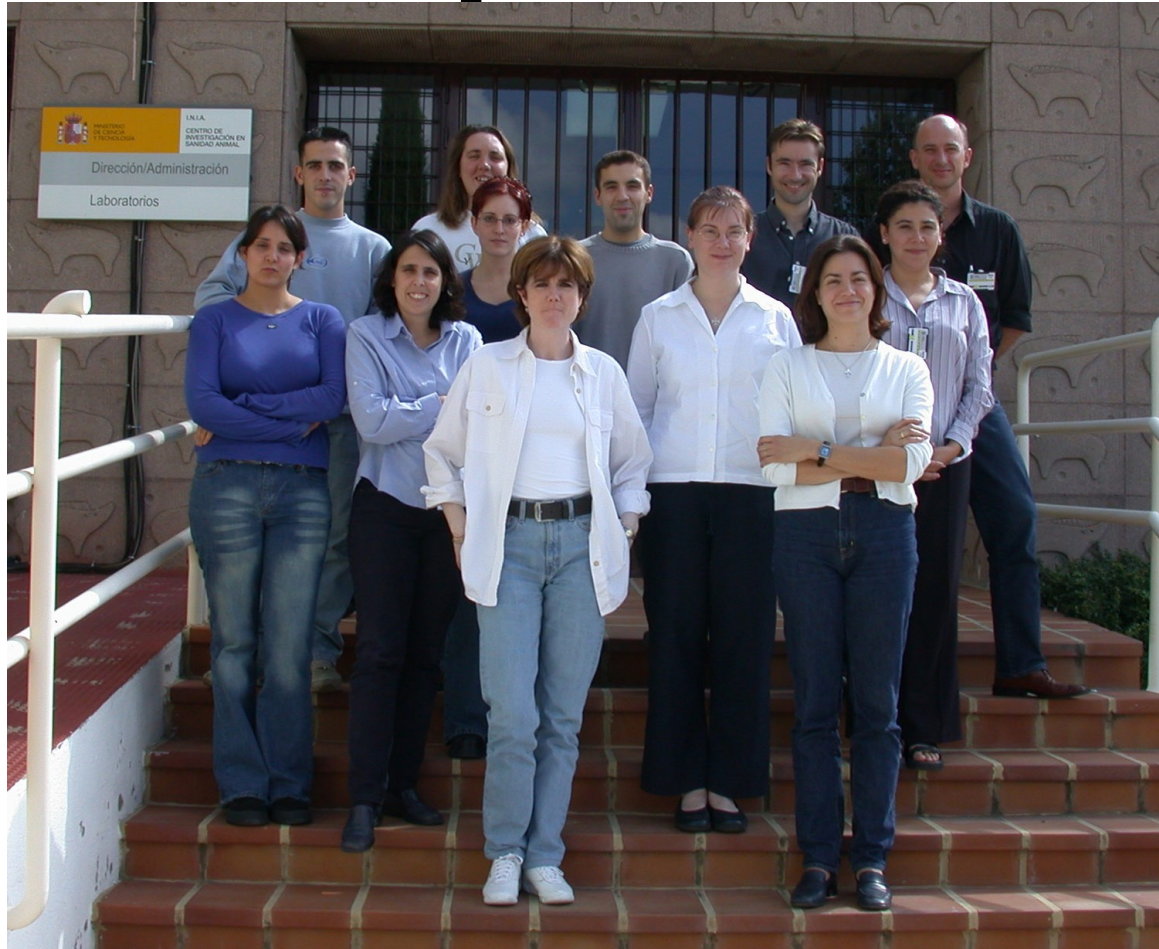
Reeves's muntjac
(*Muntiacus reevesi*)
 $2n=46$



Indian muntjac deer
Muntiacus muntjac
 $2n=6/7$

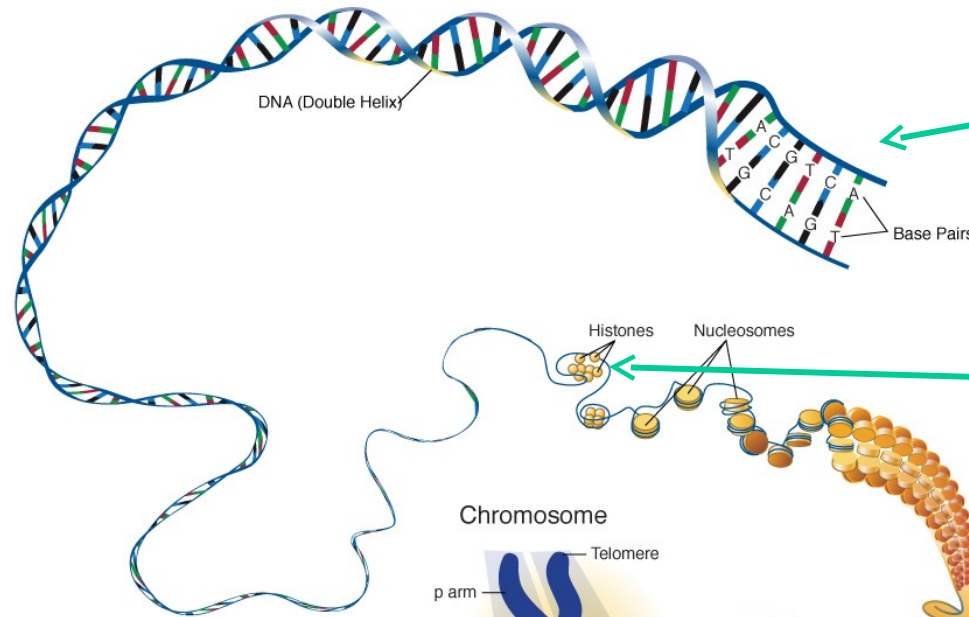


Differences within a species



Within a species there are phenotypic differences on the 'normal' spectrum. We want to understand the genomic and genetic bases of these differences

Where are the differences?



Sequence variant at base pair;
DNA modifications
(e.g. methylation)

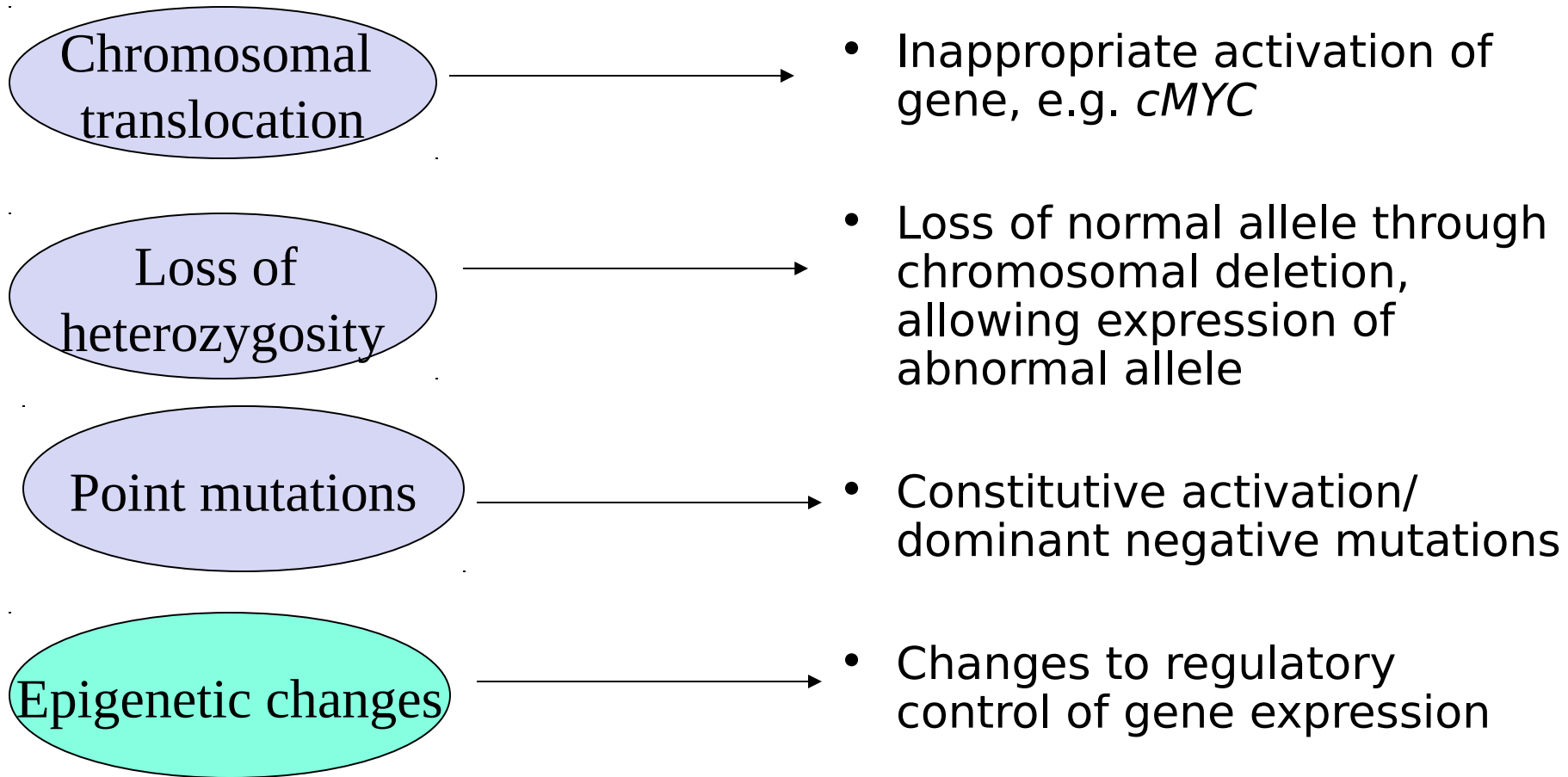
Histone modification

Chromosome:
heterochromatin
length
polymorphisms,
structural
polymorphism

DNA packaged
into
chromosomes in
nucleus; length
reduced 10,000
fold

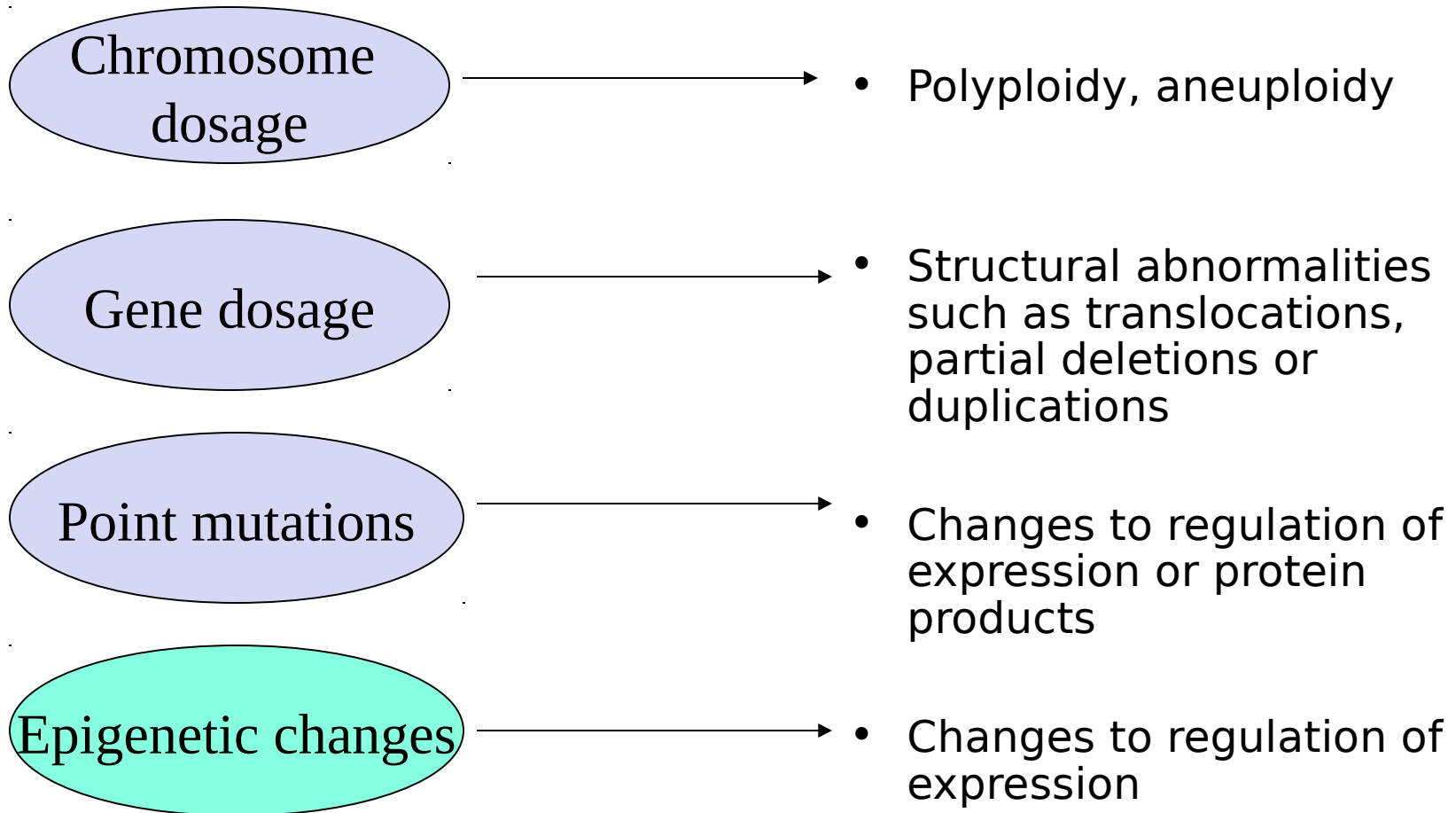
Chromosome during
DNA replication

Somatic changes



Specific examples in lectures on tumour biology (*chromosome dosage may be altered in later stages of tumour progression*)

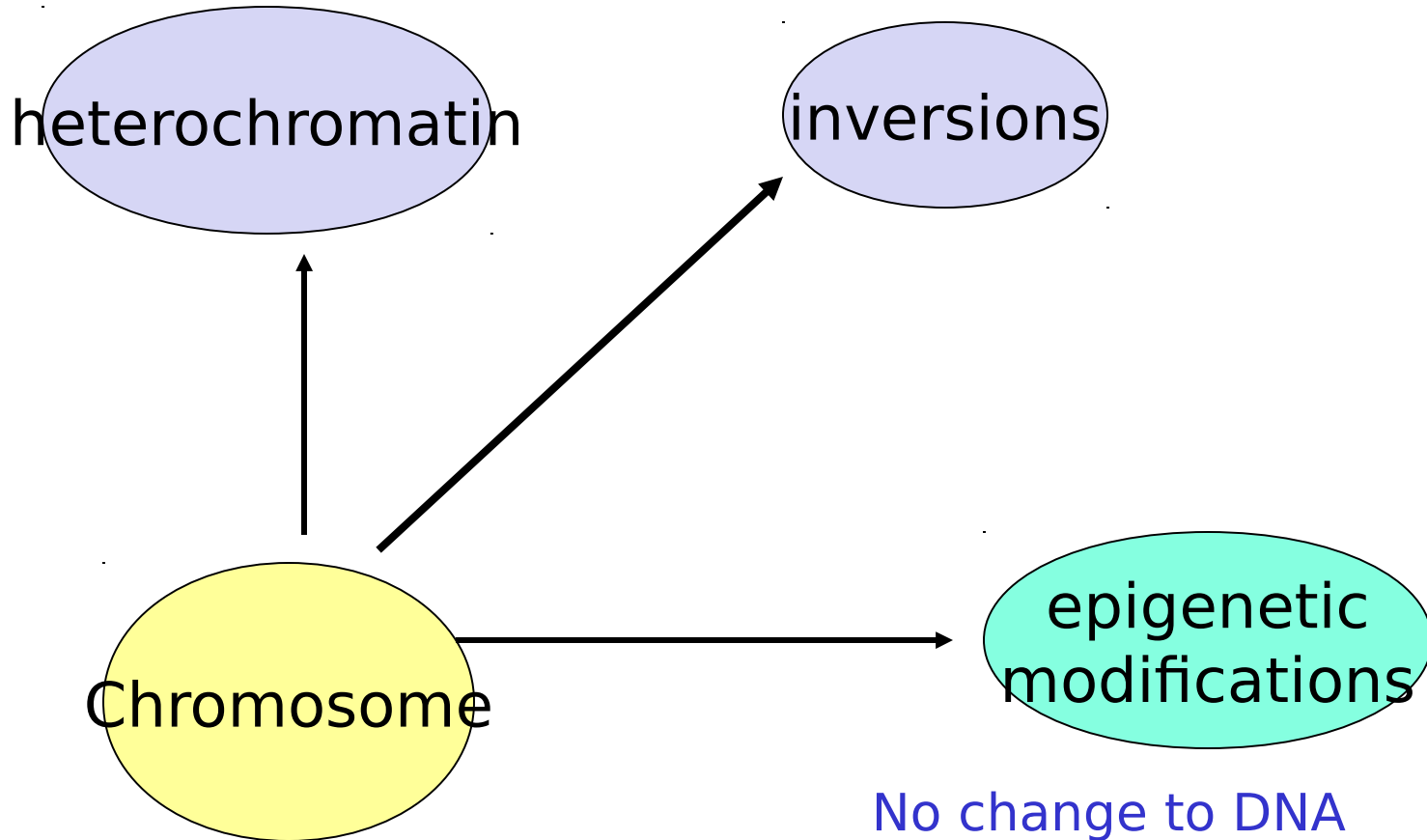
Germline changes



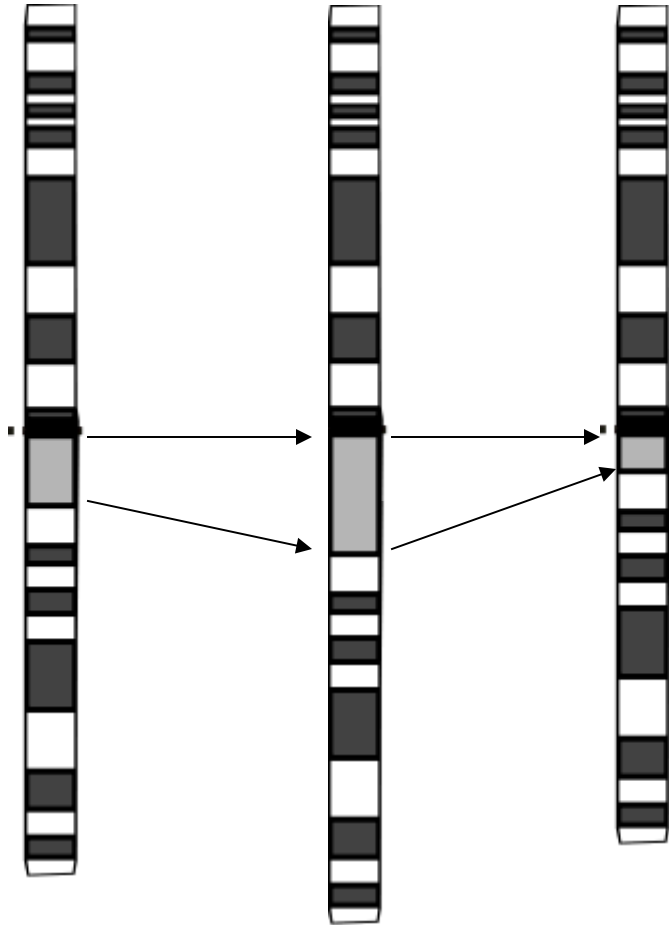
Specific epigenetic examples covered in imprinting lectures

Differences at chromosomal level

Structural and at nucleotide level. Cytogenetically visible



No change to DNA
sequence, but either bases
or histones can be modified.
Needs specialised detection
protocols



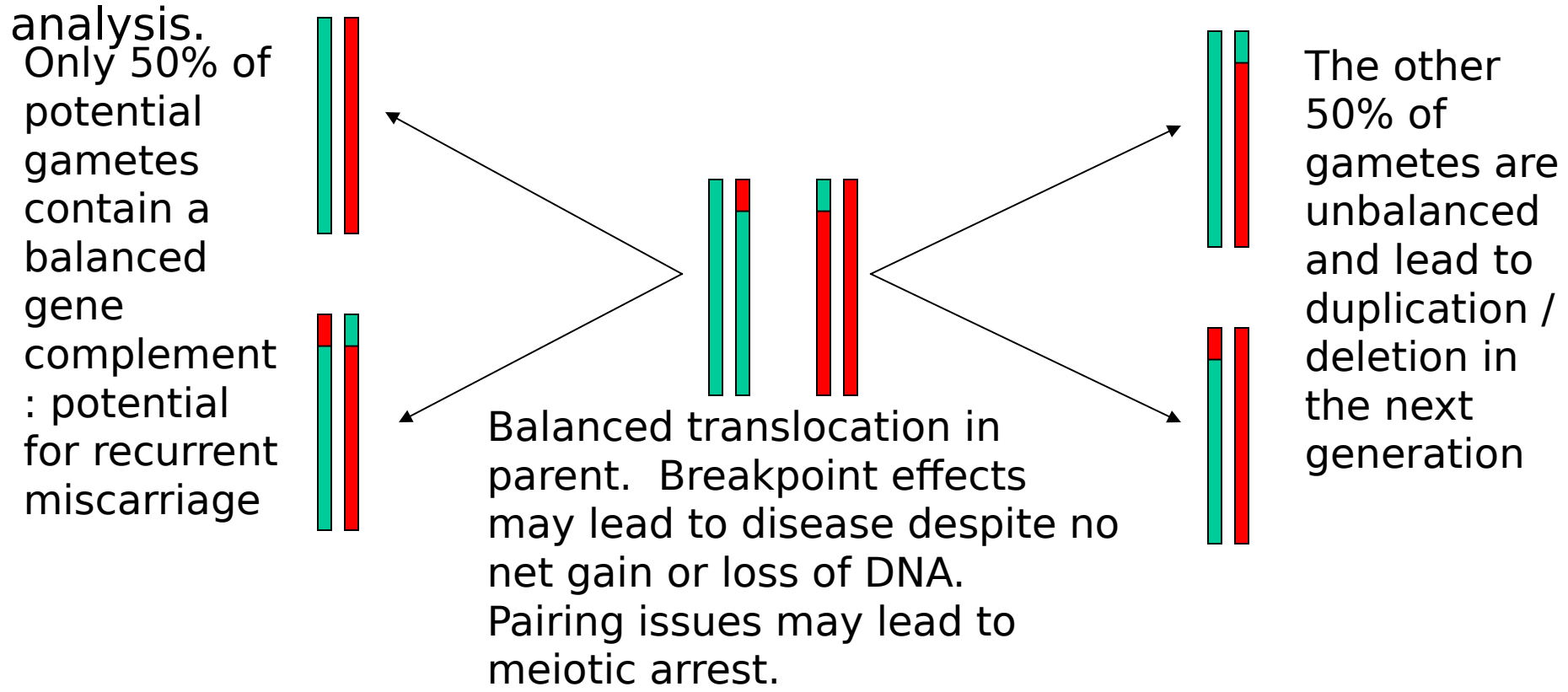
Human chromosome 1

Heterochromatin can vary in length. Can be used as a marker to cytogenetically track a chromosome through a family.

(Heterochromatin often comprises repetitive DNA sequences. Not usually transcriptionally active)

Structural abnormalities of Chromosomes

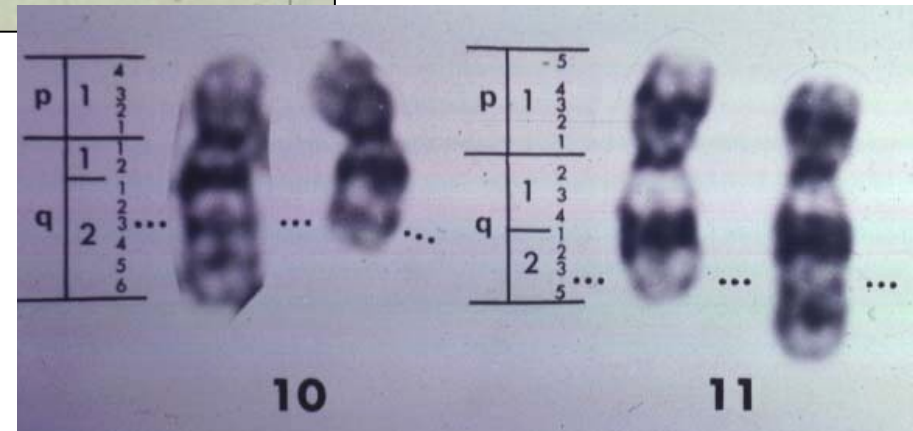
Can cause disease either by altering gene content (i.e. duplication, deletion, unbalanced translocations; disruption of genes at chromosomal breakpoints) leads to fertility problems – pairing, balanced / unbalanced translocations. Can be tracked in pedigree analysis.

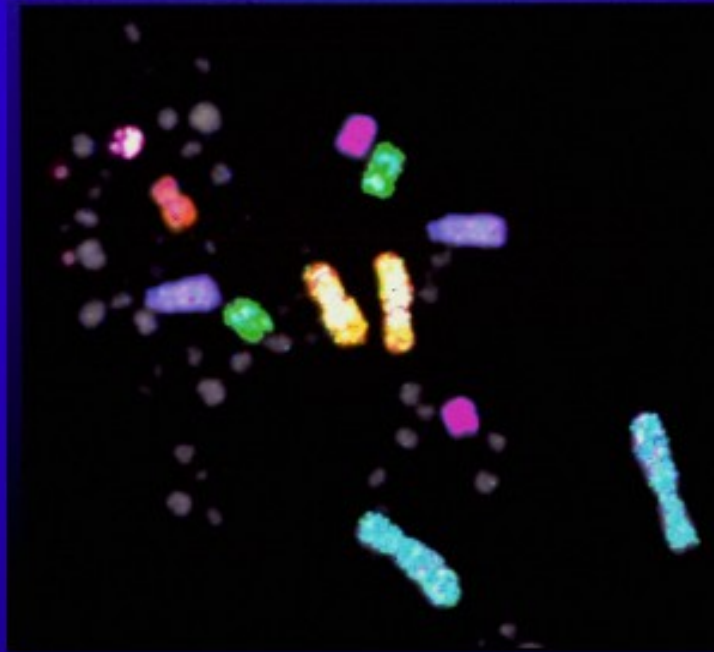




balanced translocation.

*Old data, but the translocation is obvious.
Phenotypic severity is linked to the impact on a single important gene (eg translocation in cancers) or the total number of genes involved in any unbalanced aneuloidy, NOT, the size of the translocation.*

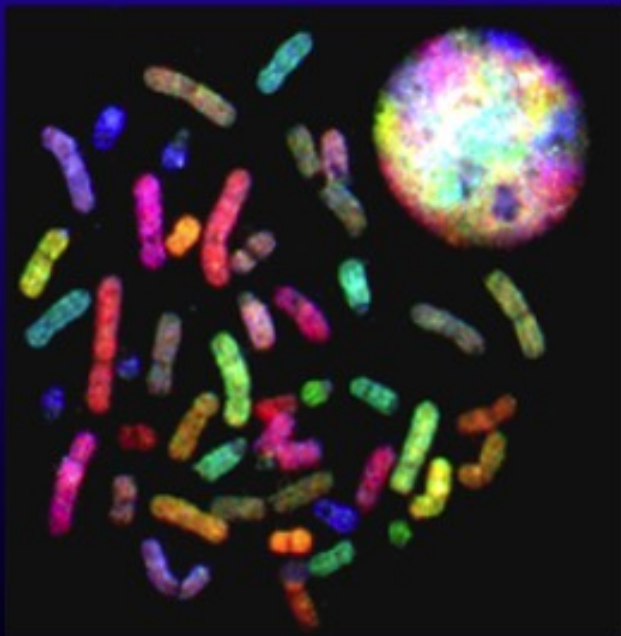




Karyotypic banding patterns

Single chromosome paints

Spectral karyotyping



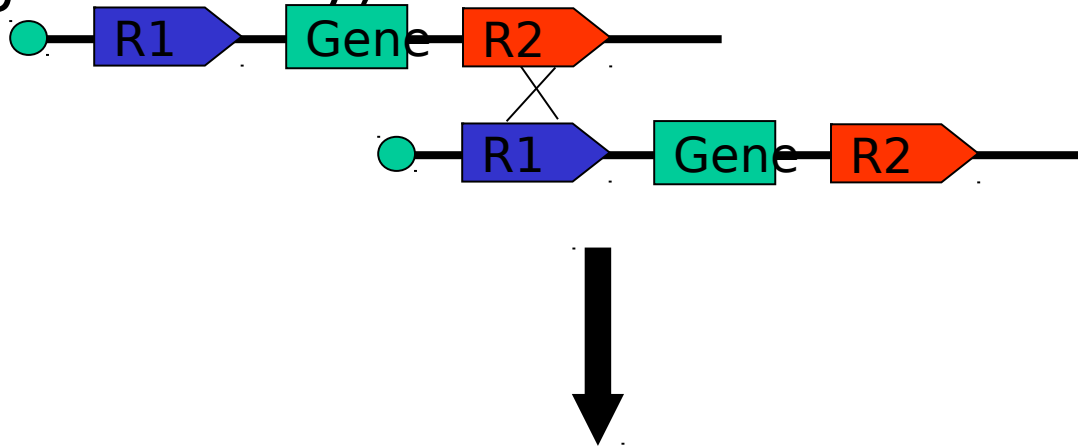
More in Prof Griffin's lectures

Chromosomal segments: Copy number variation (CNV)

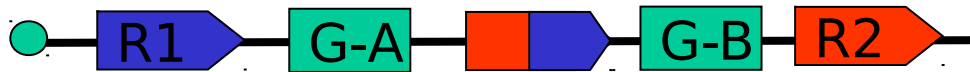
- Increasingly, identifying regions of the genome that vary in copy number between individuals.
 - Product of non-allelic homologous recombination (NAHR)
- May be up to 10% of genome that can be involved in CNVs
- *Is it just polymorphism, can this lead to disease susceptibility, or does it reflect our disease history?*

Covered in context of genome evolution, genome dynamics and disease by Drs Skinner and Hurles

NAHR: between repeats flanking genes on homologous chromosomes (could also be tandem gene family)



In this case the repeats flanking genes are involved in the recombination event

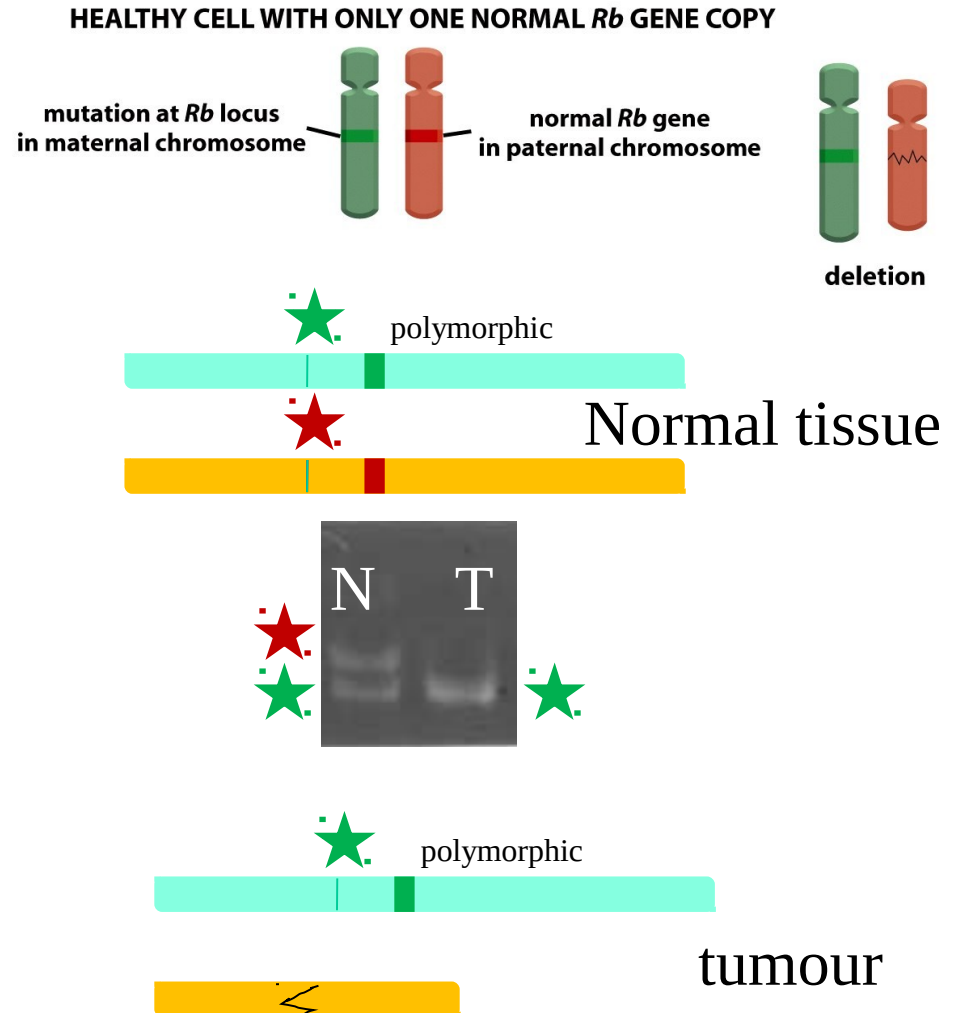


Hybrid sequence

Outcome:
duplication on one chromosome, and a reciprocal deletion on the other chromosomal product *in this example, loss of a gene*

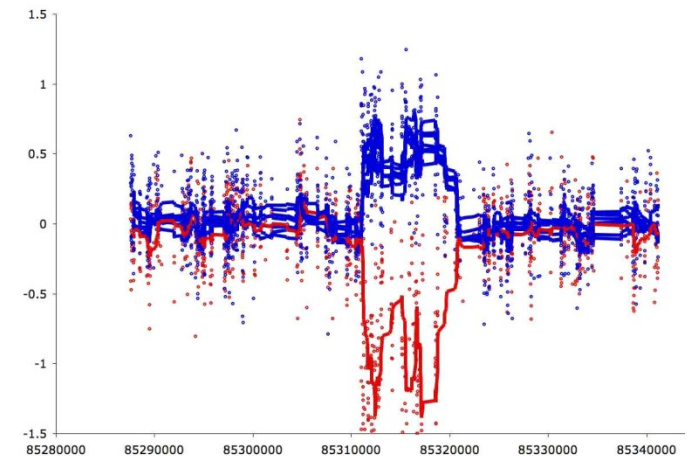
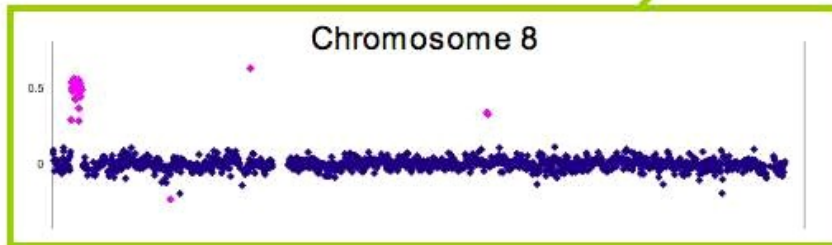
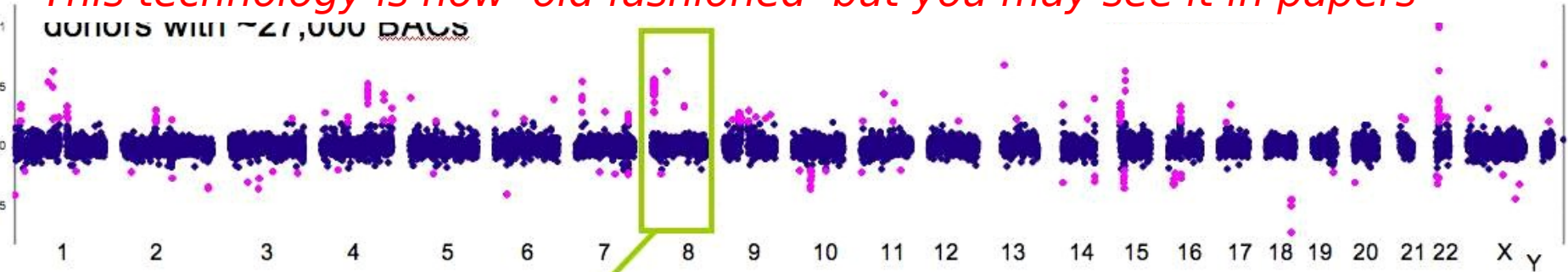
Genomic deletions and duplications are also found in cancers

- Deletion of normal allele exposes mutation in TS genes. Tumours show Loss of Heterozygosity (LOH)
- Duplications (and further amplifications) of proto-oncogenes can lead to tumour development. Identifying common genomic regions of amplification can help locate these genes.



Comparing DNA from two individuals using a microarray

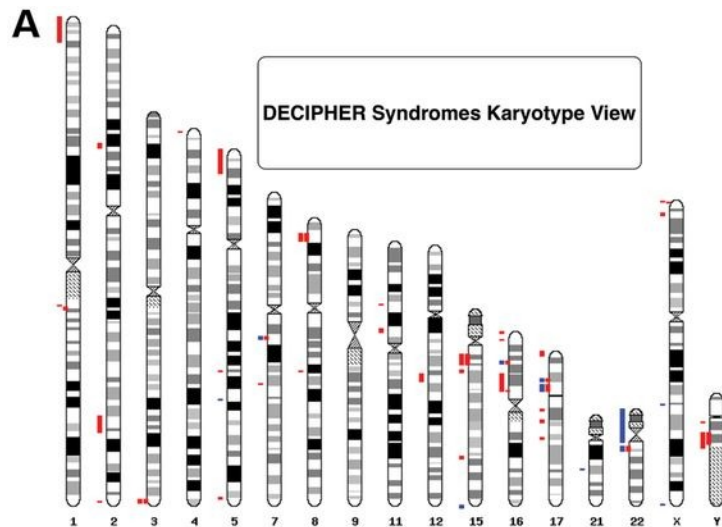
This technology is now 'old fashioned' but you may see it in papers



Expanded view of chr. 8:
High resolution showing
duplications (blue) and deletion
(red)

Adapted from Copy Number Variation website:
<http://www.sanger.ac.uk/humgen/cnv/>

Expert-reviewed syndrome resource in DECIPHER. (A) Karyotype view showing the location and nature of all DECIPHER syndromes.



This project is key to determining which CNV changes are truly pathogenic.

B 17q21.31 recurrent microdeletion syndrome

Overview Variants (1) Phenotypes (5)

Last modified: 2012-08-10 12:23:00

Clinical: The syndrome is very variable, but common features include: low birthweight (0.4th-9th centile), neonatal hypotonia, poor feeding in infancy (often requiring naso-gastric feeding for a period) and oromotor dyspraxia together with moderate developmental delay/learning disability but friendly/amiable behaviour. Other clinically important features include epilepsy, heart defects (ASD, VSD) and kidney/urological anomalies. Silvery depigmentation of strands of hair have been noted in several patients. With age there is an apparent coarsening of facial features. 17q21.3 was reported simultaneously in 2006 by three independent groups. A reciprocal microduplication of 17q21.3 had also been described (Kirchhoff 2007). Koolen (2008) provides an overview of the clinical features of the syndrome by reviewing 22 individuals with a 17q21.3 microdeletion and estimate a prevalence of ~1/16,000. Koolen (2012) and Zollino (2012) recently showed that haploinsufficiency of the gene *KANSL1*, a regulator of chromosome modification, is sufficient to cause the 17q21.31 microdeletion syndrome.

Size of deletion: The recurrent deletion is between 500-650kb in size encompassing the *KAT8* regulatory NSL complex subunit 1 (*KANSL1*) gene.

Origin of deletion: A 900-kb inversion that suppresses recombination between ancestral H1 and H2 haplotypes encompasses the deletion. The orientation of LCRs flanking the deleted segment in inversion heterozygotes, or H2 homozygotes, sponsors this microdeletion by means of non-allelic homologous recombination (NAHR) and Koolen (2008) showed that in the 5 cases examined, the parent originating the deletion carried the common 900kb 17q21.31 inversion polymorphism (H2 haplotype) $p < 10^{-5}$.

Expert advisor: Dr Helen V Firth, Consultant Clinical Geneticist, Addenbrooke's Hospital, Cambridge, UK

Links to support groups:

www.rarechromo.org

Links to further information:

www.geneclinics.org

www.orpha.net

17q21.com




C 22q11 deletion syndrome (Velocardiofacial / DiGeorge syndrome)

[Print Report](#)

Overview Variants (1) Phenotypes (5) Citations (8) Karyotype

1 to 1 of 1 variants

	Location	Interval (Mb)	Copy Number	Genes	UCSC/et
22:19009792-21452445		2.44	0	43	

10 per page

10 per page

Graph Genes (43) Patient overlap (129) Syndrome overlap (2)

All (43) OMIM (37) Morbid (7)

1 to 10 of 43 entries

Name	Location	Description	OMIM	Morbid	%HI 2	Links
TBX1	22:19744226-19771116	T-box 1	✓	✓	26.6	O E I
SNAP29	22:21213271-21245502	synaptosomal-associated protein, 29kDa	✓	✓	97.9	O E I
SERPIND1	22:21128167-21142008	serpin peptidase inhibitor, clade D (heparin cofactor), member 1	✓	✓	10.4	O E I

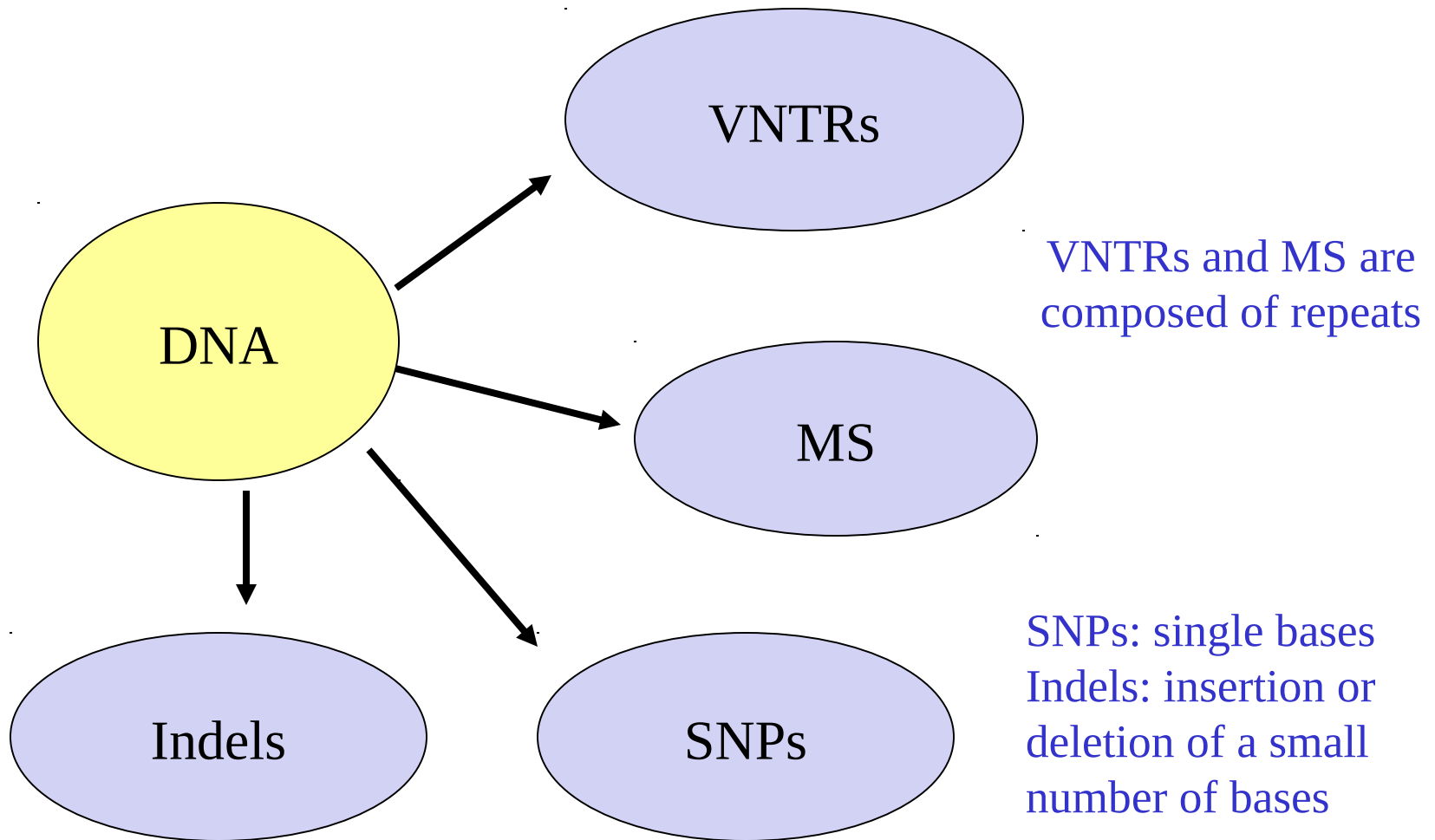
Swaminathan G J et al. Hum. Mol. Genet. 2012;21:R37-R44

Properties of genetic markers

1. Genetic marker: ideally has a known location in the genome, must be polymorphic in the population being studied, must be easy to assay ([NOTE THAT PROTEINS HAVE BEEN USED PRE-DNA](#))
2. markers are most useful if they have high information content (i.e. lots of alleles at good frequencies)
 - high heterozygosity level in population
 - » For biallelic systems as close to 0.5 as possible, 0.05 for minor allele frequency (MAF) as minimum
 - It also helps if we can predict the parental origins of the alleles in offspring for pedigree analyses

[Lectures on pedigree and population analysis \(Affara and Sargent\)](#)

Variation at DNA sequence level

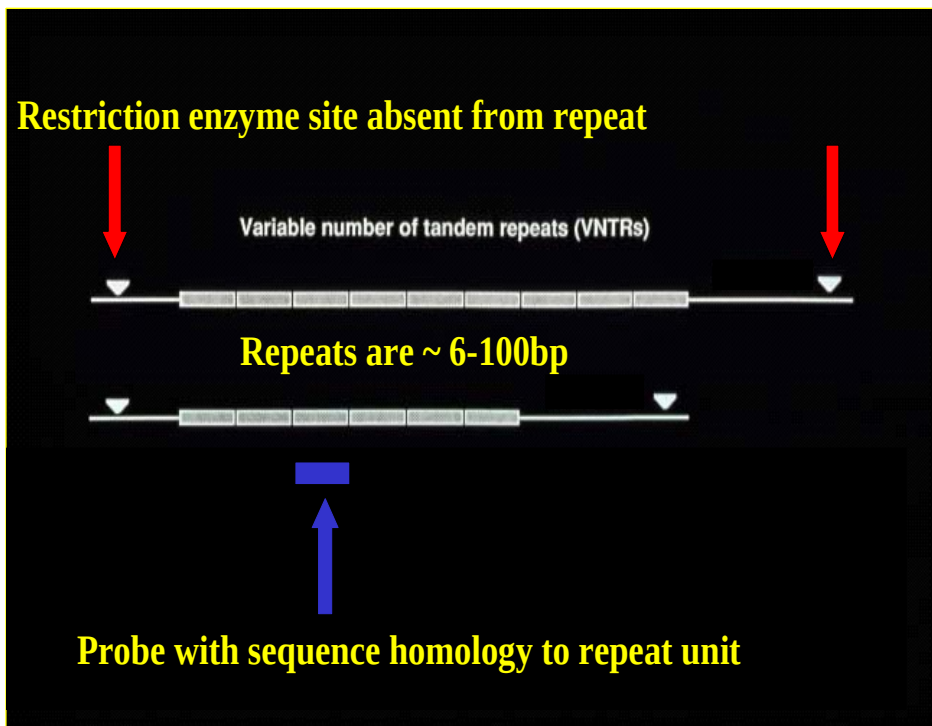


Variation in DNA sequence occurs as a consequence of mutation

- Mutations in DNA sequence arise owing to errors in DNA replication or recombination at cell division
- **Somatic** cell or in **germline (*heritable*)**
- Estimated 2×10^{-7} chance of mutation per gene per cell division
- Most DNA variants are polymorphisms, but some may be pathogenic
 - Impact on the phenotype may determine whether a change becomes established in a population
 - Positive selection, negative selection, null

Repetitive DNA- Minisatellites

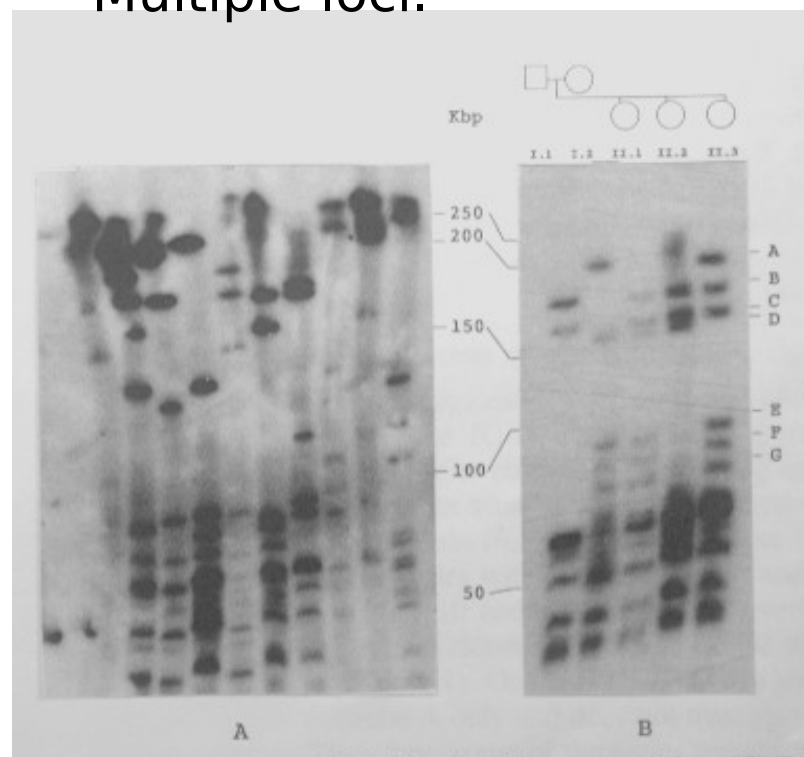
- Variable number tandem repeats (VNTRs)
 - 6-100bp per repeat unit
 - total length of multiple tandem units may be many Kb
 - core sequences in these repeats may occur scattered over genome, i.e. they are multilocus as well as multiallelic
 - VNTRs were the original basis of fingerprinting techniques for forensics and other types of DNA testing.
 - NOT EVENLY DISTRIBUTED IN GENOME (genome architecture lecture)



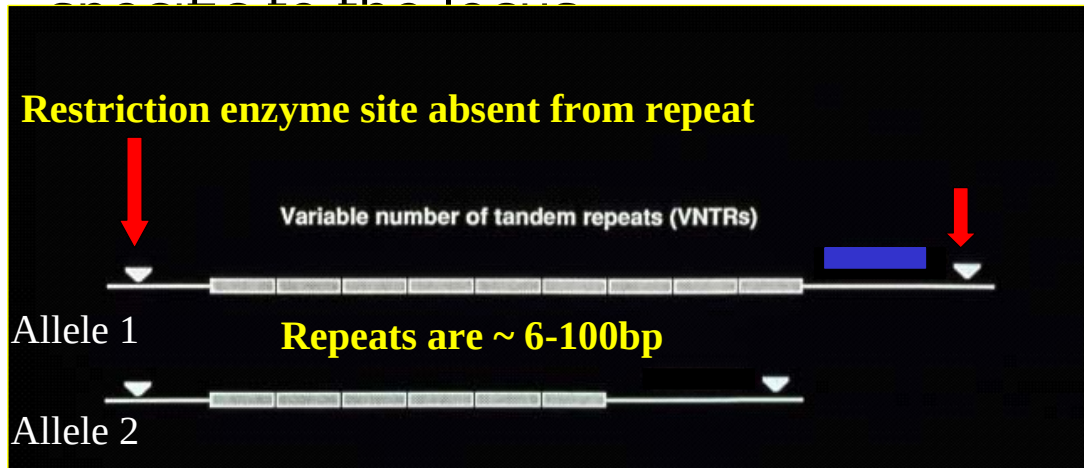
Probes designed from the repeat sequence can pick up related sequences at multiple loci, and on different chromosomes. **Complex DNA fingerprints, BUT, you can still trace specific bands in the pedigree.**

Many bands.
Complex pattern.
Forensic fingerprinting.

Downside- needs a lot of DNA.
Not locus specific.
Multiple alleles.
Multiple loci.



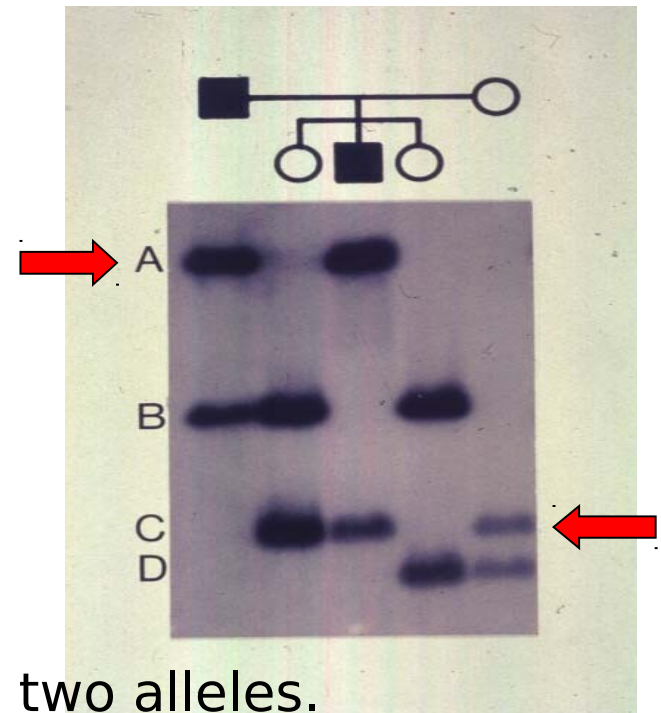
Probes designed from the unique sequence adjacent to the repeat is



Alleles of different size due to variable numbers of repeats. Each individual has two alleles.

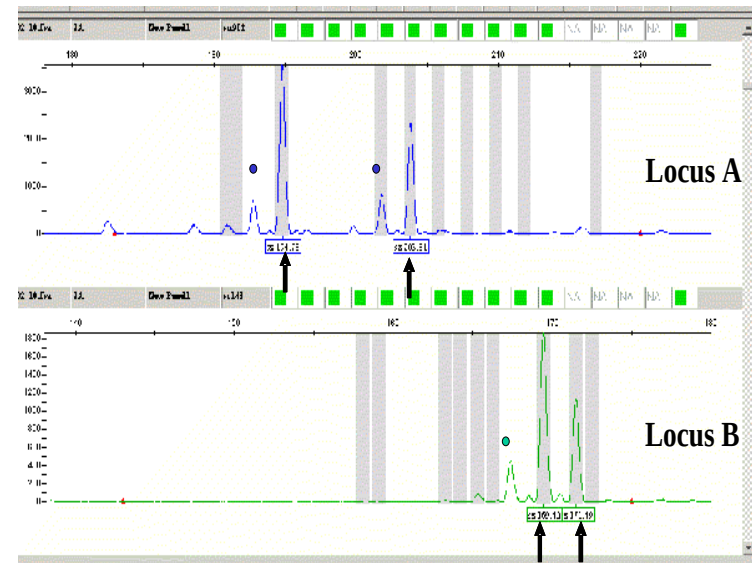
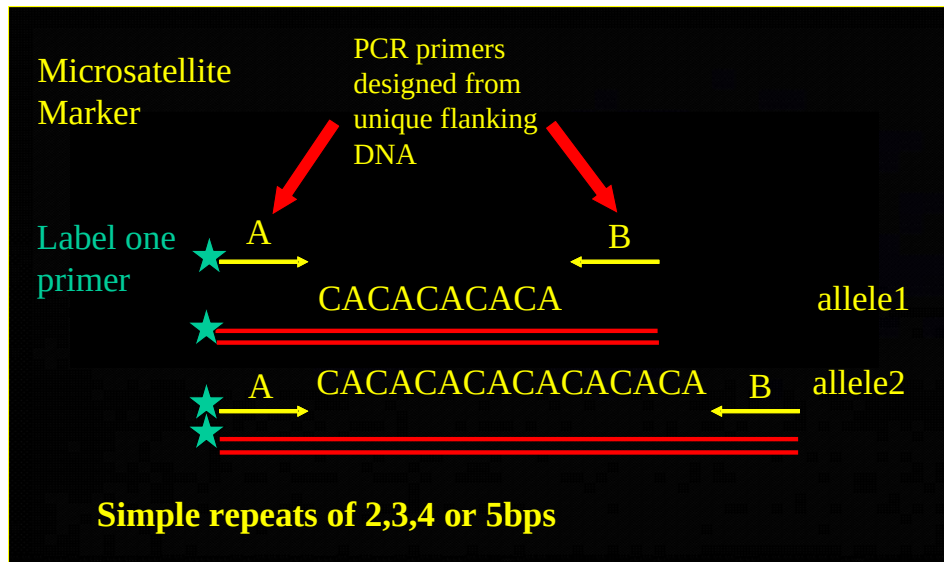
In this example, each allele inherited by the son can be assigned a parental origin. In this pedigree where father and son have a phenotype not shared by the sisters, allele A is a co-inherited marker for the phenotype.

Still needs a lot of DNA BUT is locus specific.



Microsatellites (MS)

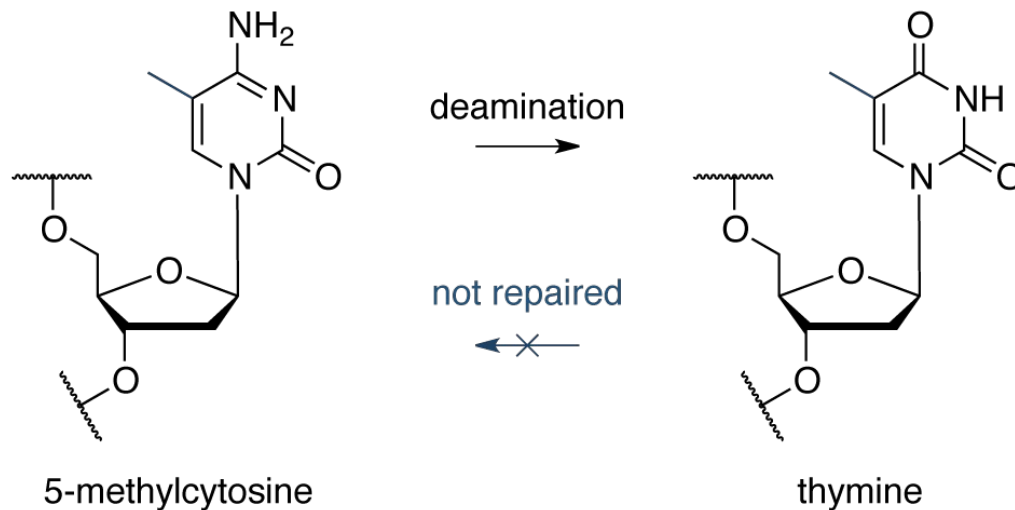
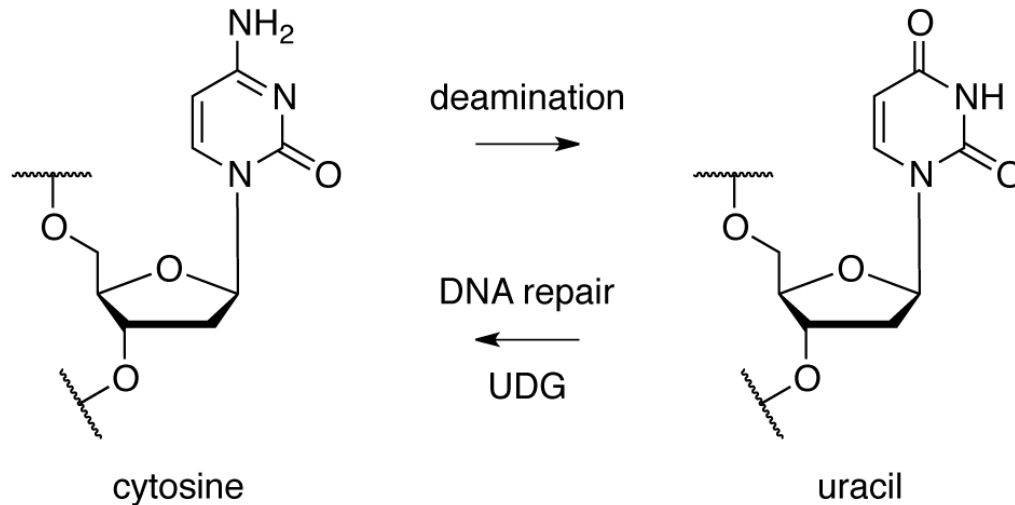
- Simple sequence repeats (SSR), short tandem repeats (STR)
- Usually 1-5bp per repeat unit
- Very frequent in genome (at least one per 100Kb), and more evenly distributed
- Can easily develop single locus assays that detect the two alleles, shorter overall length, so more amenable to PCR based analysis



Using different fluorescent tags also allows analysis of multiple markers in a single PCR reaction. Alleles separated by size on capillary sequencers. Smaller peaks due to replication slippage

Single Nucleotide Polymorphisms (SNPs)

- Individual base changes in the DNA sequence
- Frequent (at least 1/1000bp)
- Large sequencing efforts to define all common variants in different populations: HapMap and 1,000 genomes project
- Usually biallelic
 - Need to improve power of detection by using knowledge of haplotypes and linkage disequilibrium
- Most SNPs will be purely polymorphic, but some are functionally important
 - functional SNPs
 - can be in different regions of the gene
 - alter expression of transcript
 - alter coding potential of protein product

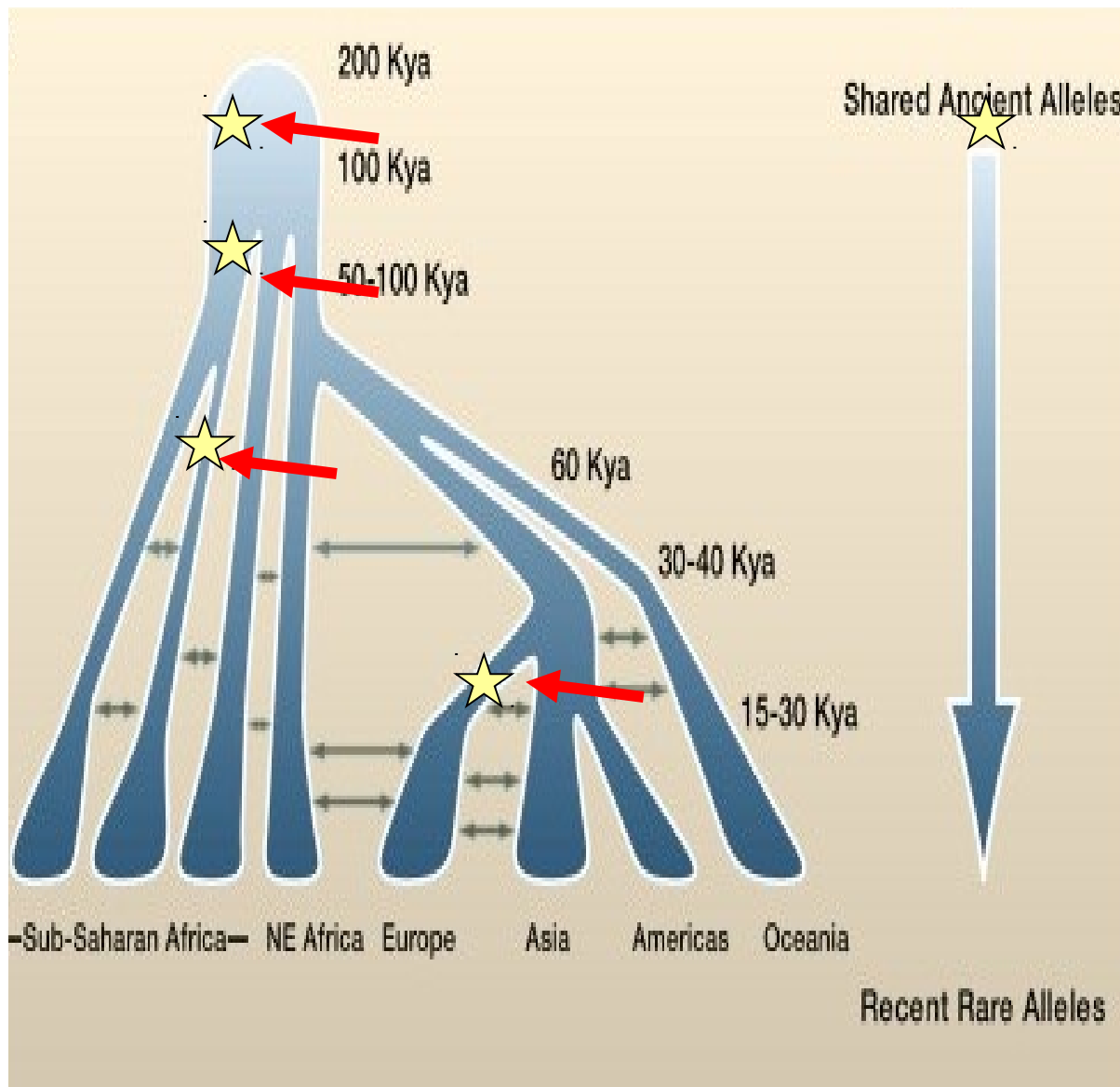


Copyright © ATDBio Ltd. 2005-2012

The commonest base changes observed when looking for SNPs are $\text{C} \rightarrow \text{T}$ and $\text{G} \rightarrow \text{A}$.

This happens through deamination. $\text{C} \rightarrow \text{U}$ is easily repaired in DNA, as U is not usually used (RNA only). However, 5-methyl C is deaminated to T, which means DNA repair mechanisms cannot always recognise the correct base from the error.

5mC often found in the context of CpG dinucleotides: these are underrepresented in the genome, and



New polymorphisms arise all the time. The oldest will be found in all populations. Others will be population specific, and the rarest may be pedigree specific, or individual specific.

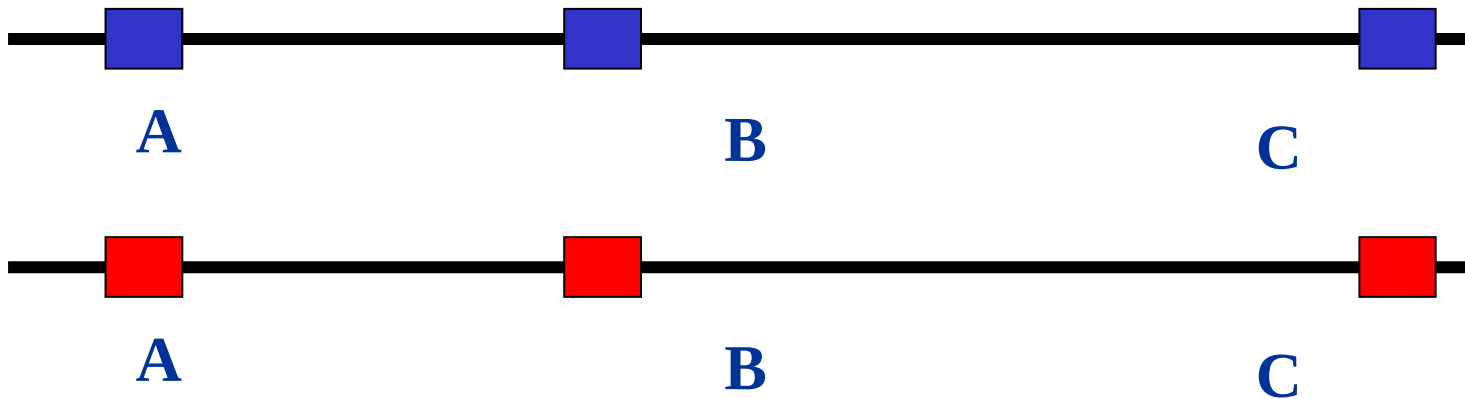
Each time a branch arises through human migration, the new population goes through a 'bottleneck' reducing genetic diversity.

Adapted from McClellan J, King MC. Cell. 2010 141:210-7.

Some Useful Terms

- Locus:
 - Specific region of the genome
- Polymorphism
 - Variation at a locus
- Allele:
 - variant at a locus
 - alleles can be discussed at level of an individual or population
- Linkage disequilibrium (LD):
 - specific combinations of alleles at numerous loci (*see next few slides*)
 - reflects evolutionary history within a population
 - LD blocks are defined stretches of genomic sequence, often flanked by recombination hotspots
- Linkage
 - within a pedigree between disease and locus. Will find recombinants over time. Allele at locus varies between pedigrees (underpinned by LD in a single pedigree).
- Association
 - reflects LD across a population with a functional variant (specific allele). Will change slowly over time.

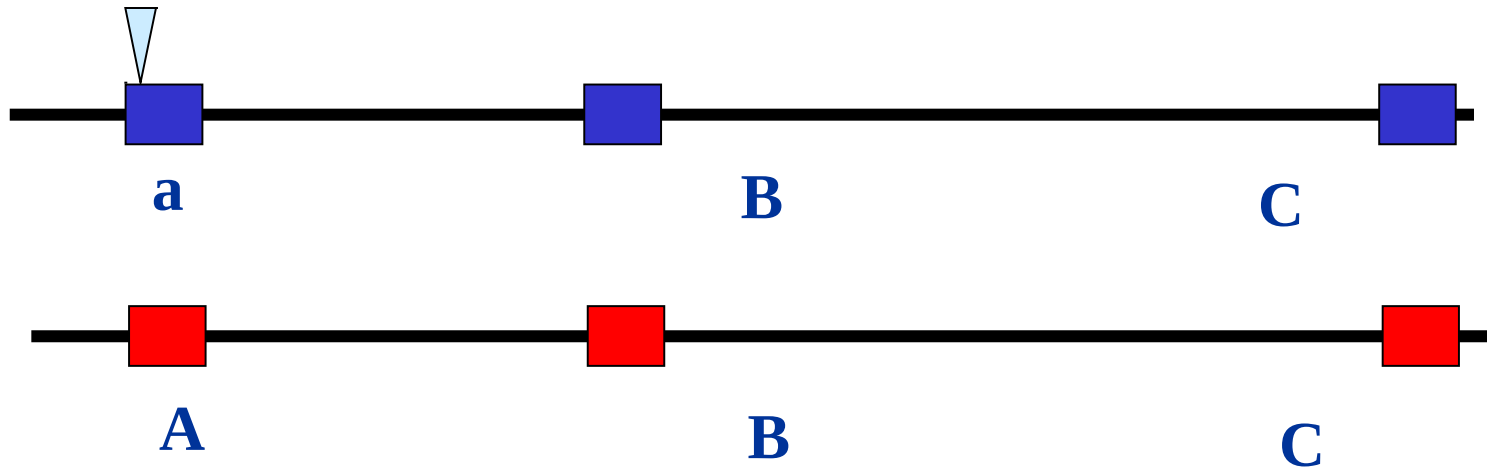
Why? SNPs, Linkage Disequilibrium (LD) and Haplotypes



Change 1: at position A

SNPs, LD and Haplotypes

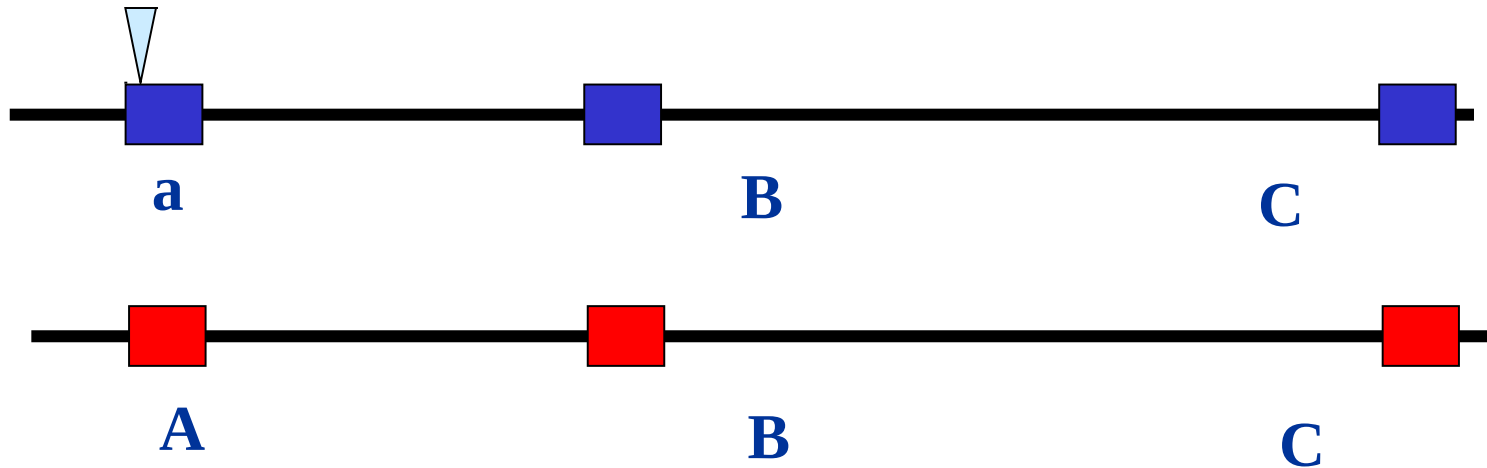
 SNP



Change 1: creates chromosomes defined as ABC or aBC. Transmission through germline to next generation.

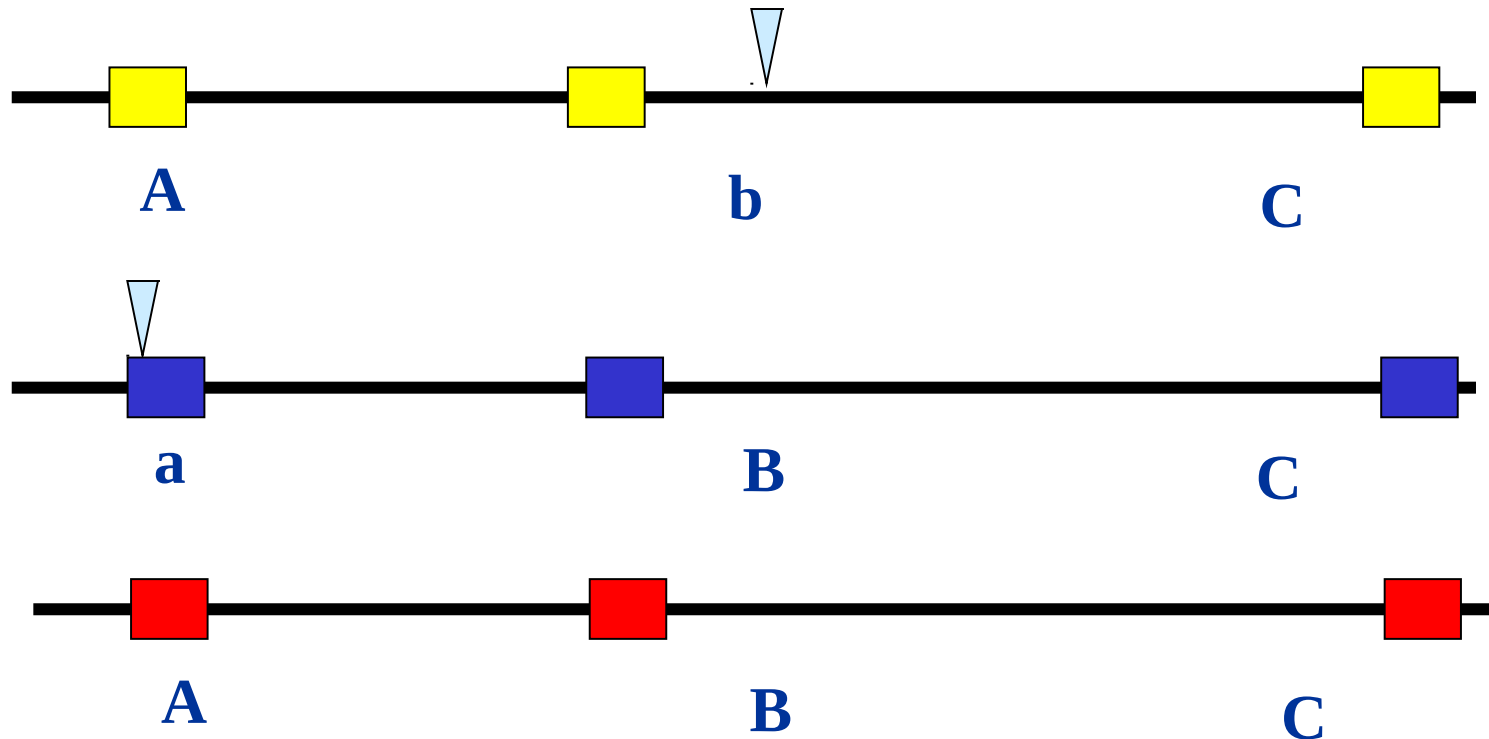
SNPs, LD and Haplotypes

 SNP



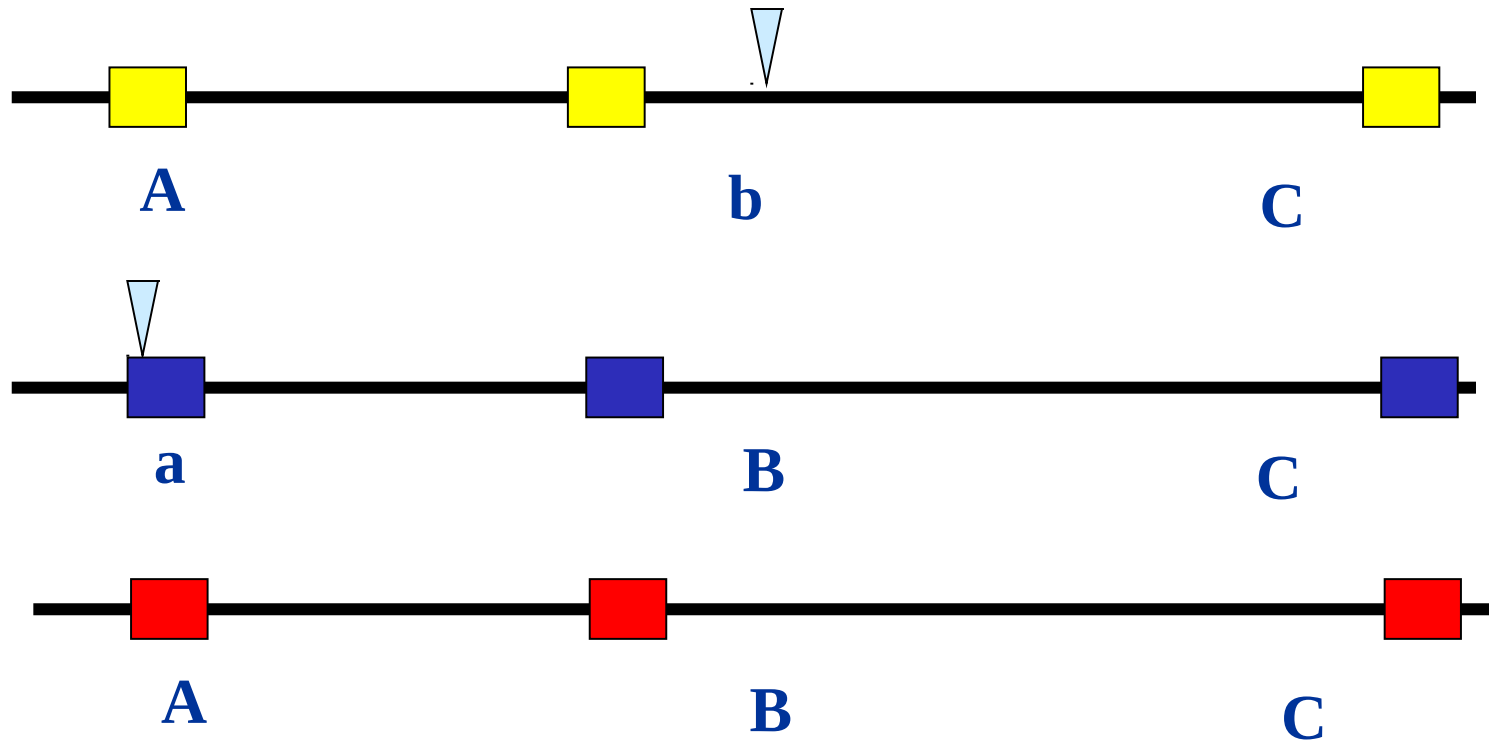
Change 2: can occur on only one of the two chromosomes now present in the population, e.g. at B on ABC

SNPs, LD and Haplotypes



Change 2: three chromosomes in the population: AbC, aBC, ABC. The combinations of markers establish the *haplotypes* along the chromosome.

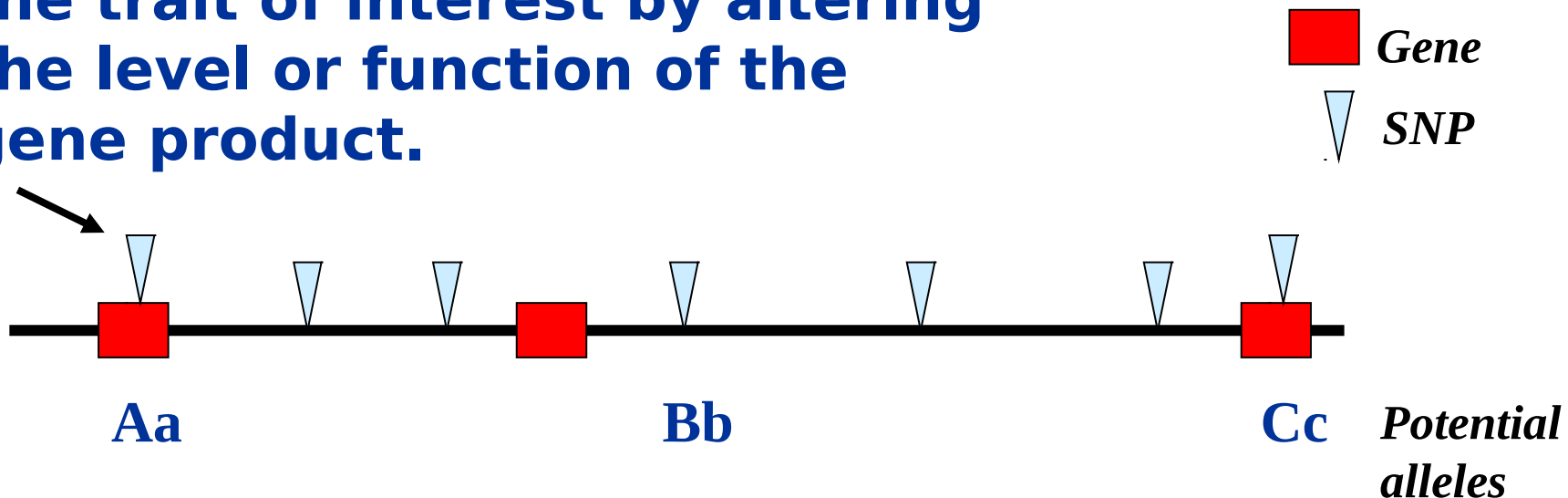
SNPs, LD and Haplotypes



The haplotypes present at the **population level** are used to define *linkage disequilibrium* blocks. LD is eroded over time via recombination, gene conversion, new mutation etc. - reflection of the evolutionary history of the population.

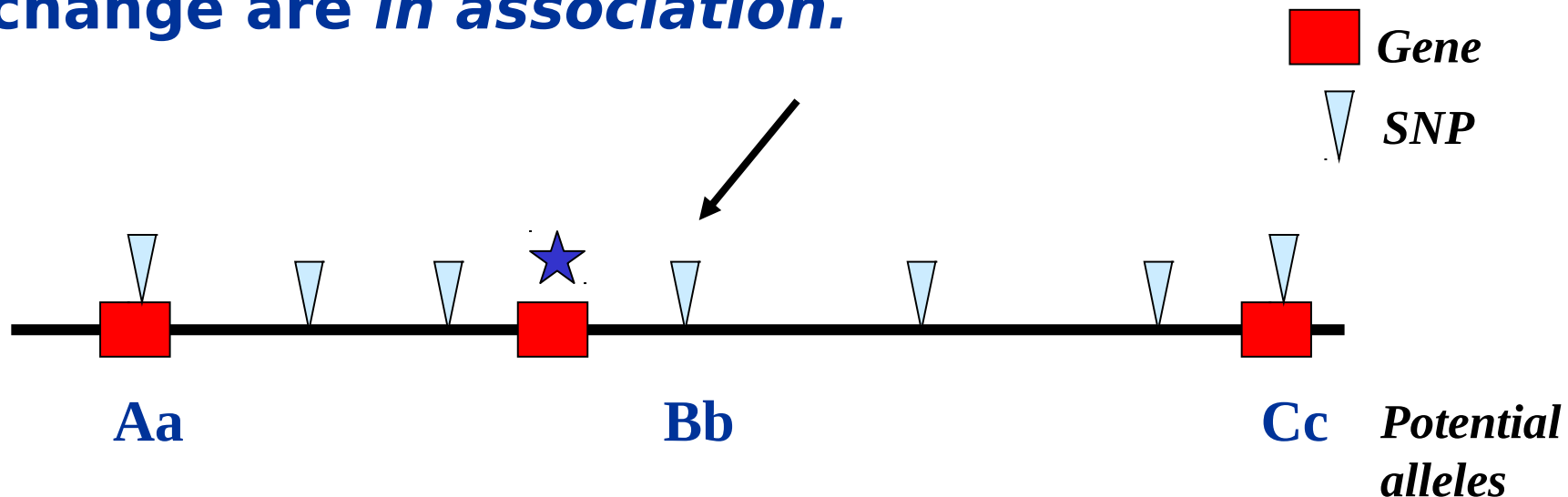
Using LD and Haplotypes

A SNP in a gene or control region could be *causative* for the trait of interest by altering the level or function of the gene product.

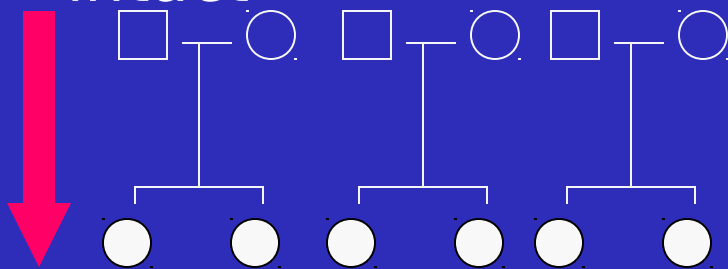


Using LD and Haplotypes

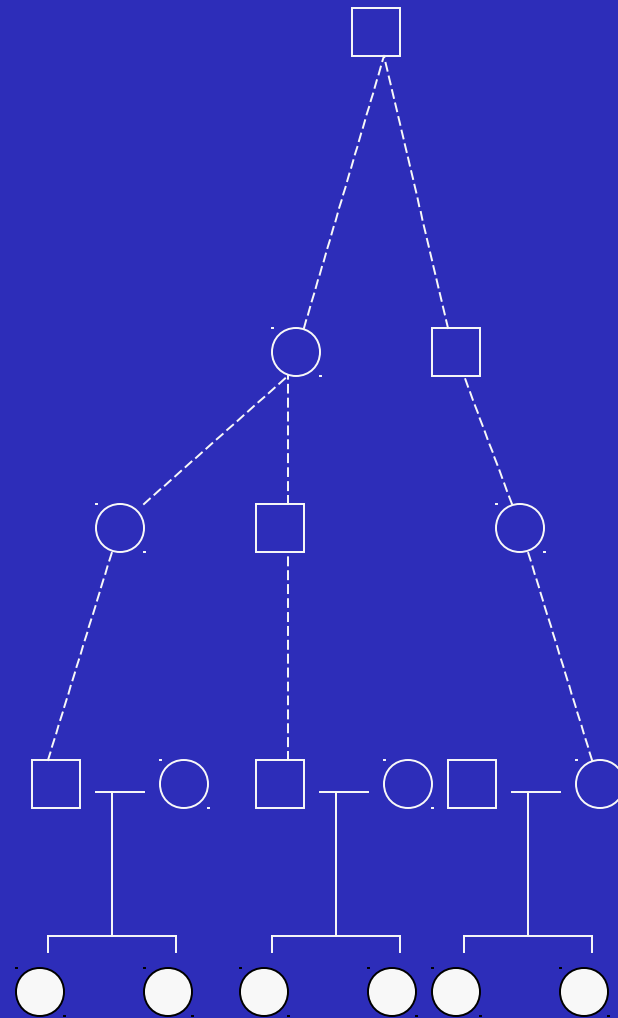
SNPs nearby that are in *linkage disequilibrium* with the causative change are *in association*.



Combinations of alleles along a section of DNA will be coinherited, but over time recombination will reduce the length of the segment of the DNA that remains 'intact'

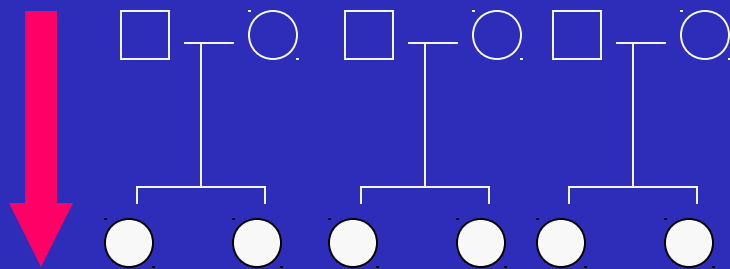


Linkage analysis
pedigrees

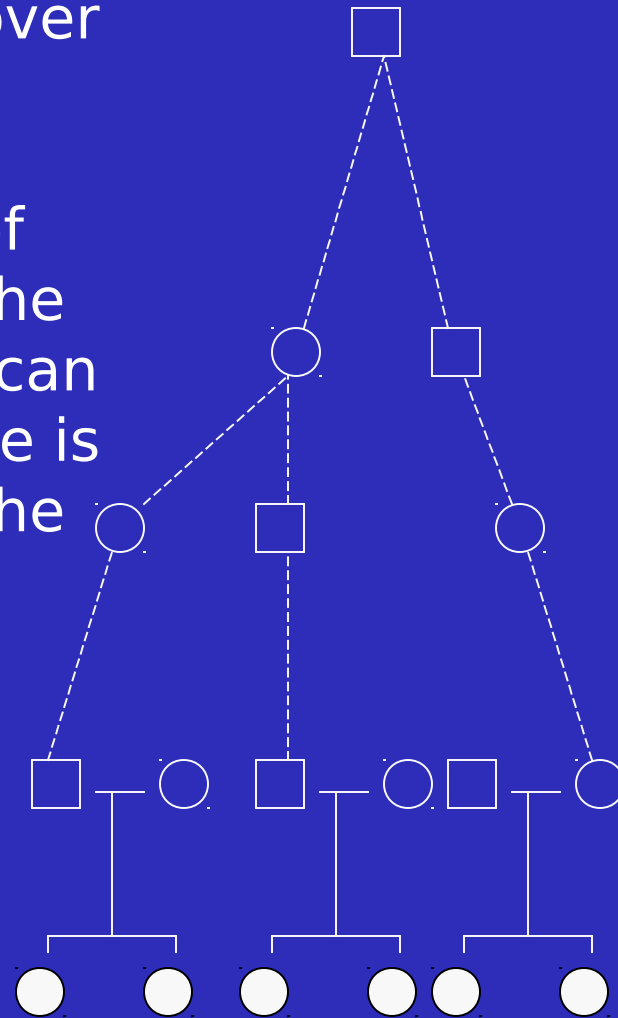


LD mapping
populations

Linkage in pedigrees occurs over a large physical portion of a chromosome . In different pedigrees the combinations of alleles are not identical, but the segment of the genome that can be tracked with the phenotype is similar. **Linkage** is always to the **Locus**



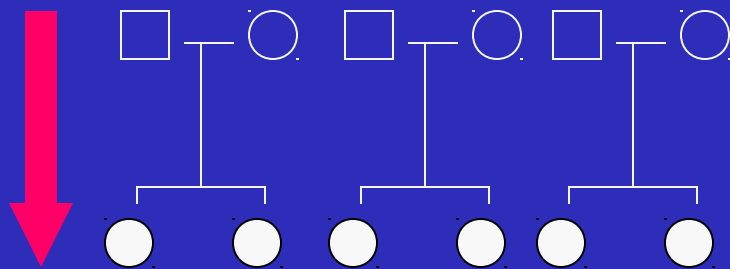
Linkage analysis



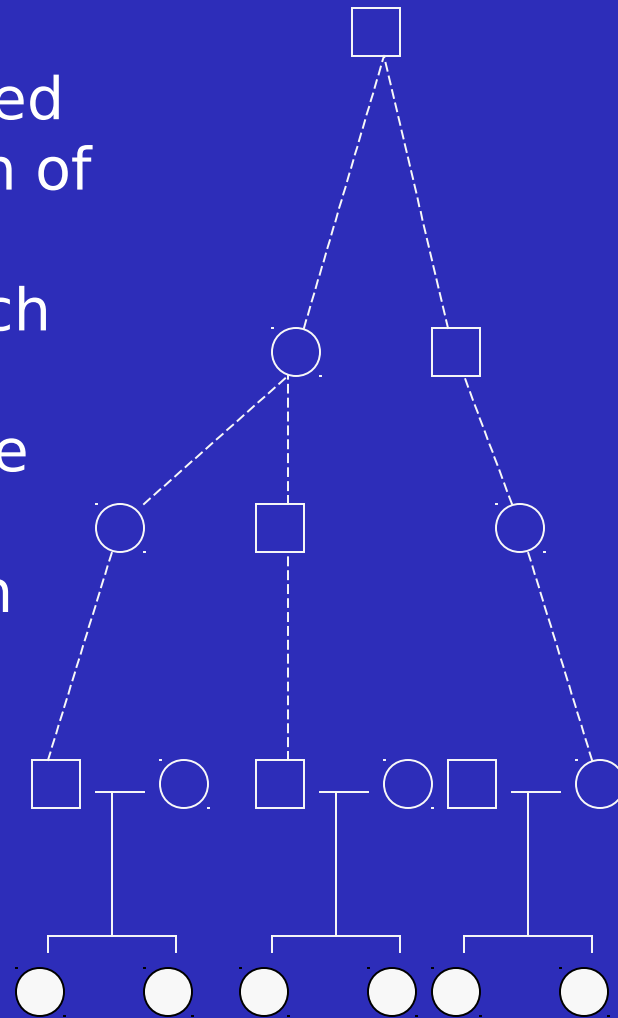
LD mapping

At the population level, sequence variants that contribute to a complex phenotype are assumed to be older, the common region of the DNA sequence shared by affected individuals will be much smaller, and the alleles should reflect the sequence around the original mutated locus.

Association is not just based on position but also on **allele**

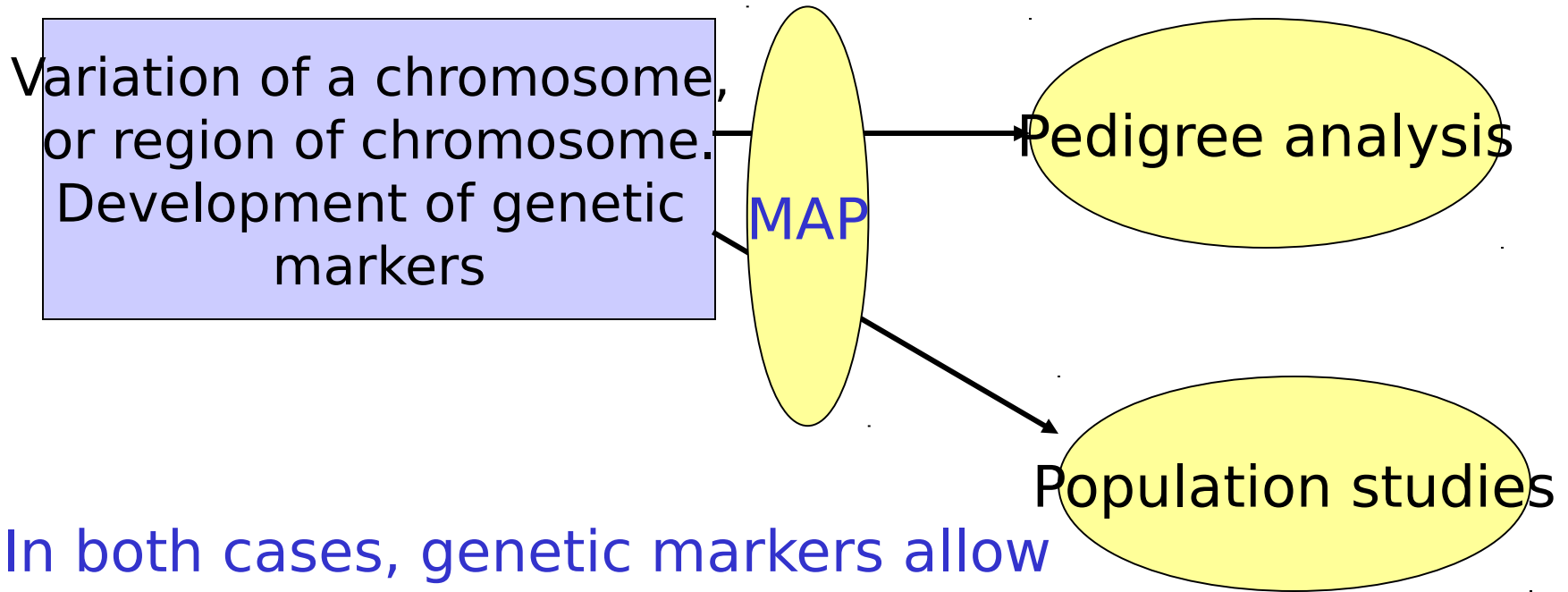


Linkage analysis



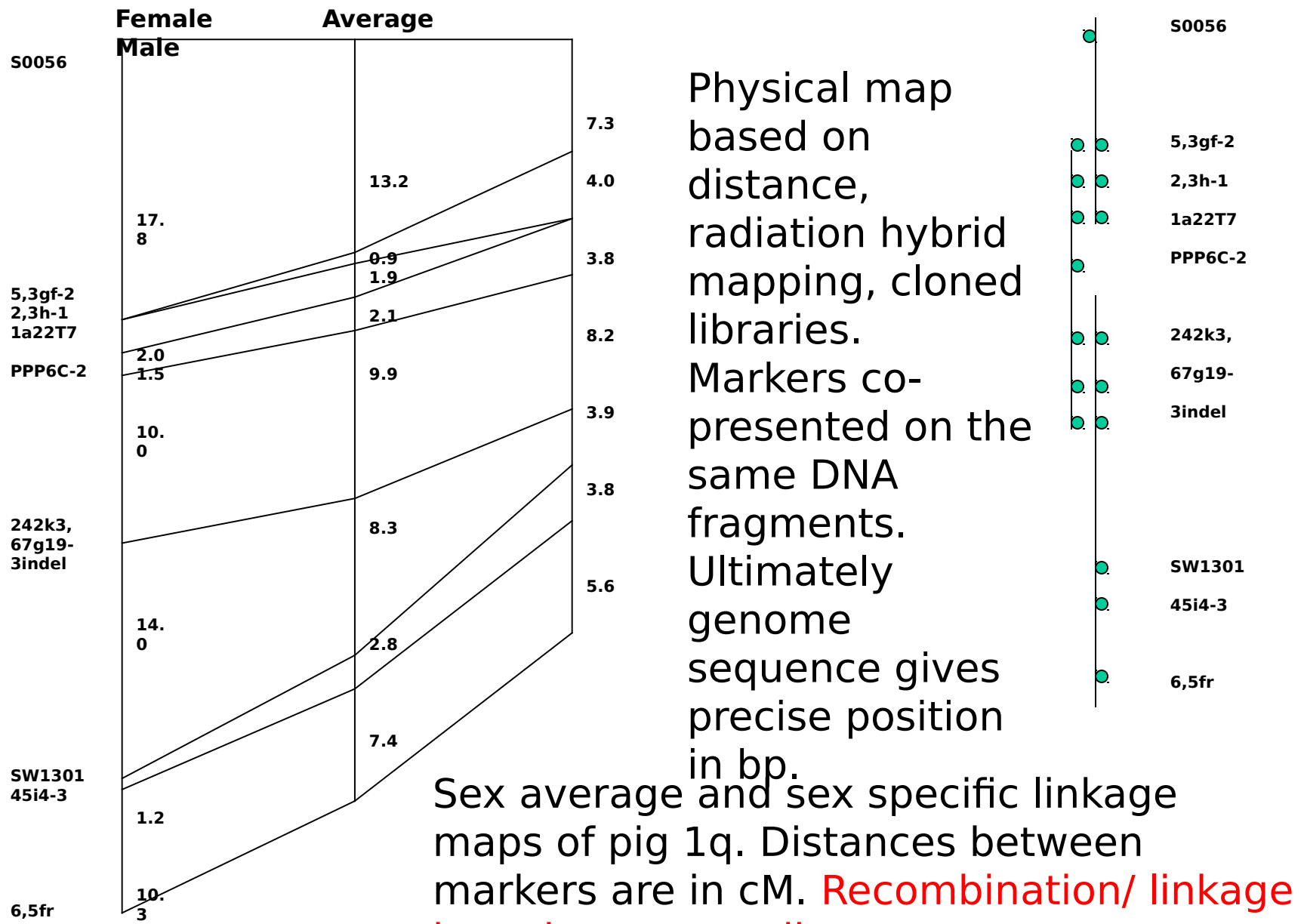
LD mapping

Using variation in disease locus mapping

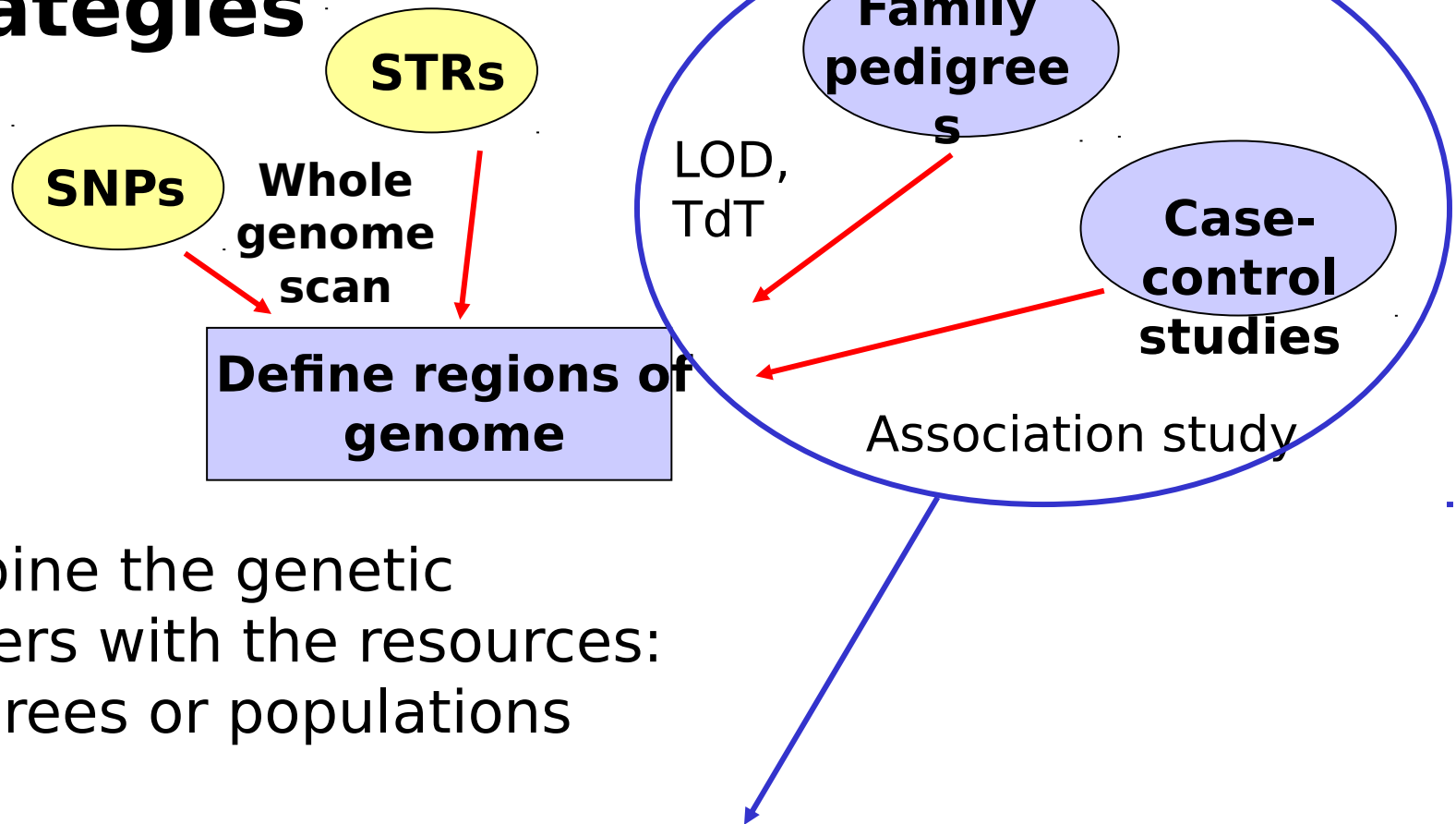


In both cases, genetic markers allow us to track regions of chromosomes that harbour genes that contribute to the phenotypic trait of interest. (This includes disease causing mutations.)

Genetic and Physical maps



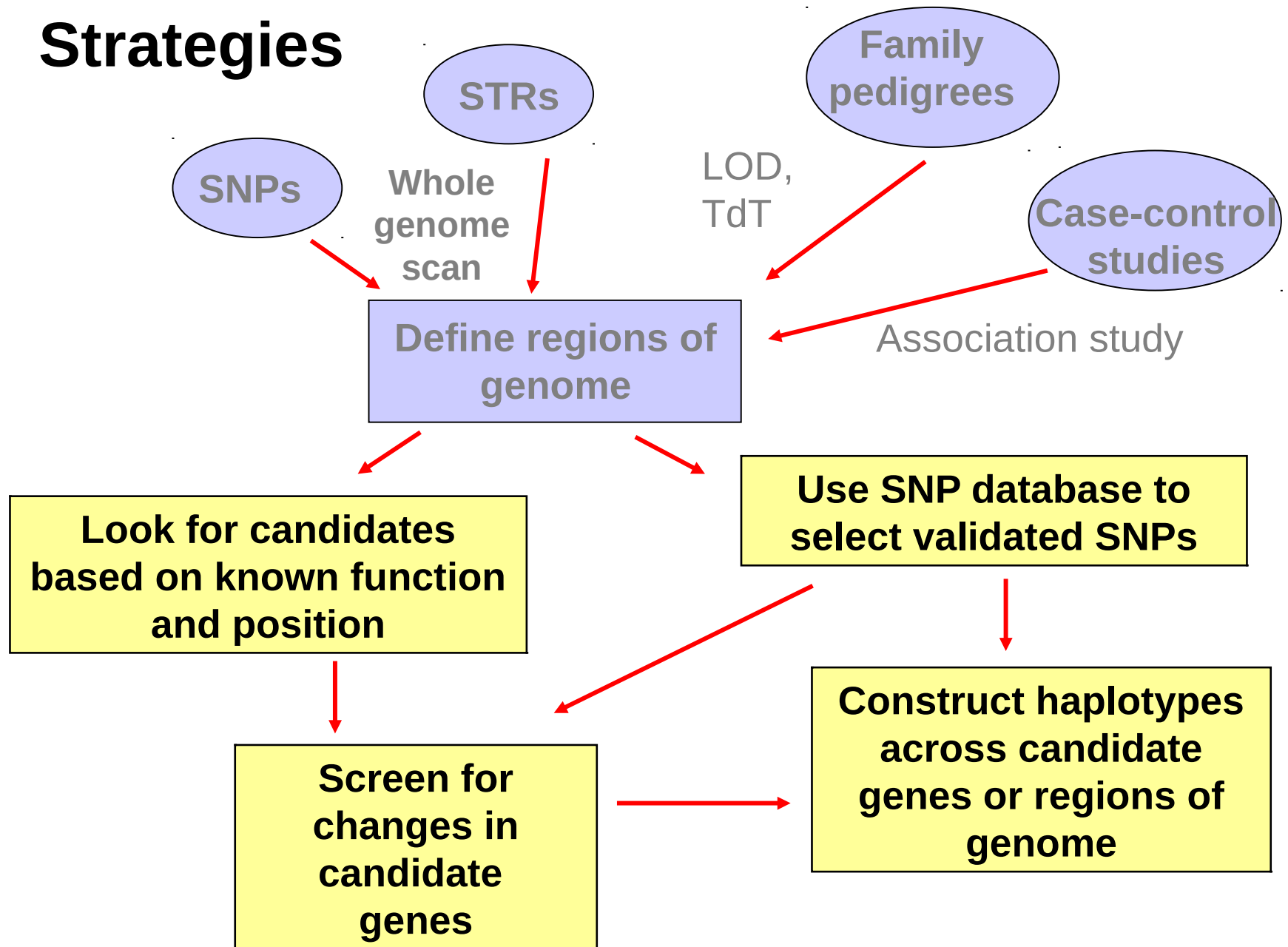
Strategies



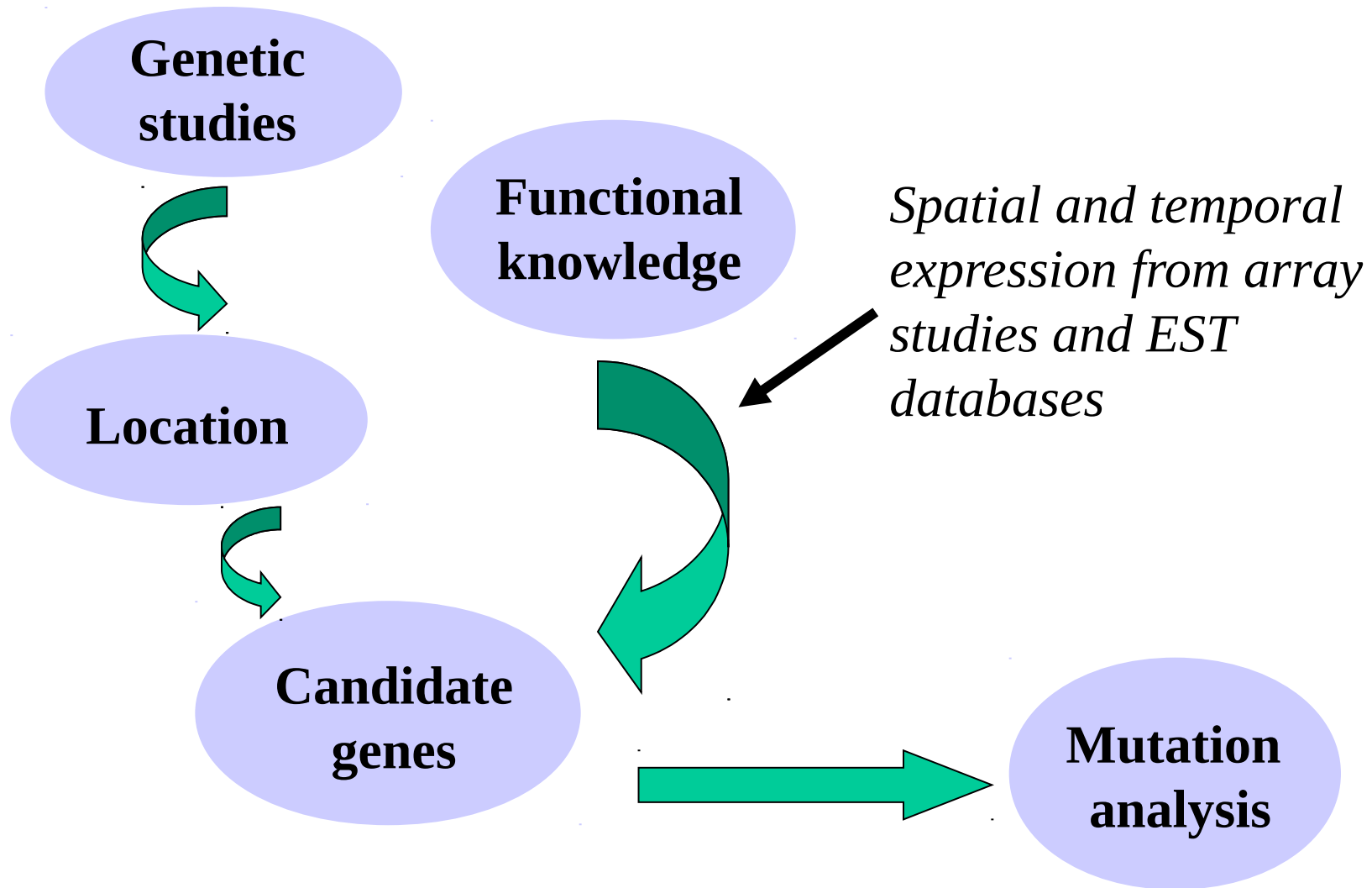
Combine the genetic markers with the resources: pedigrees or populations

Pedigree analysis models covered in more detail in Prof Affara's lectures, and more about genome wide association studies later this year.

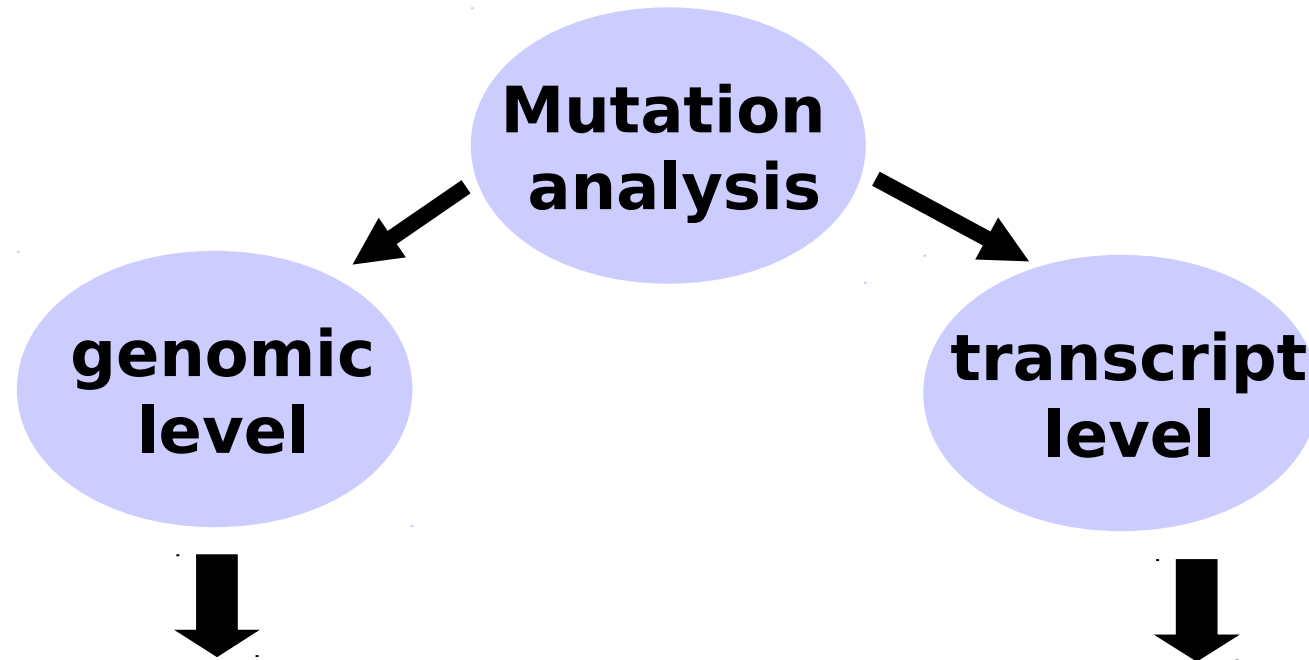
Strategies



Selecting candidates



Analysing candidates



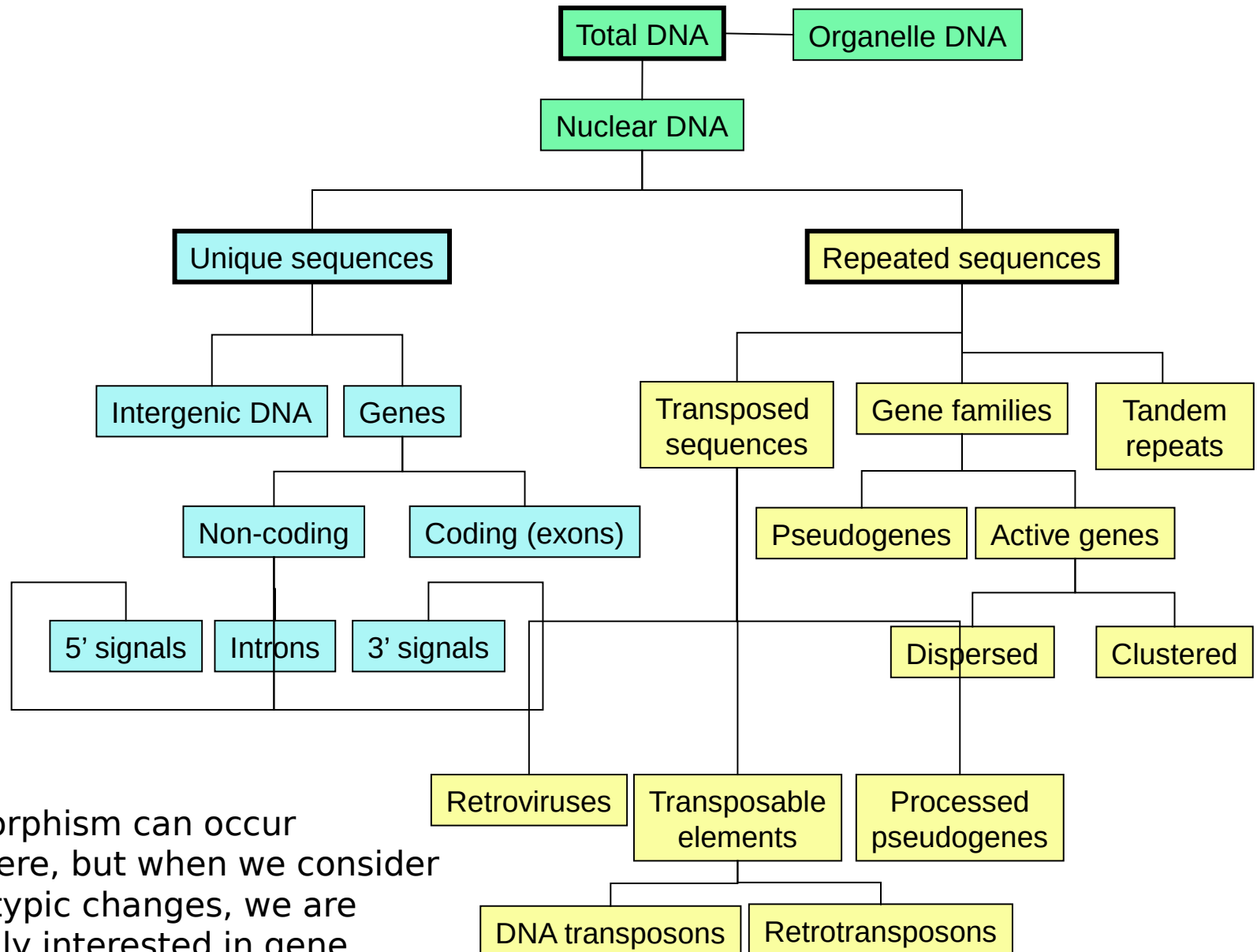
Screen each exon, splice junctions, plus promoters and other control regions.

Can detect changes missed at cDNA level.

New resequencing methods

Use cDNA to look for abnormal splicing variants, screen ORF and UTRs. May miss mutations that result in rapid degradation of abnormal mRNA

Resequencing for mutation analysis: what do we want?



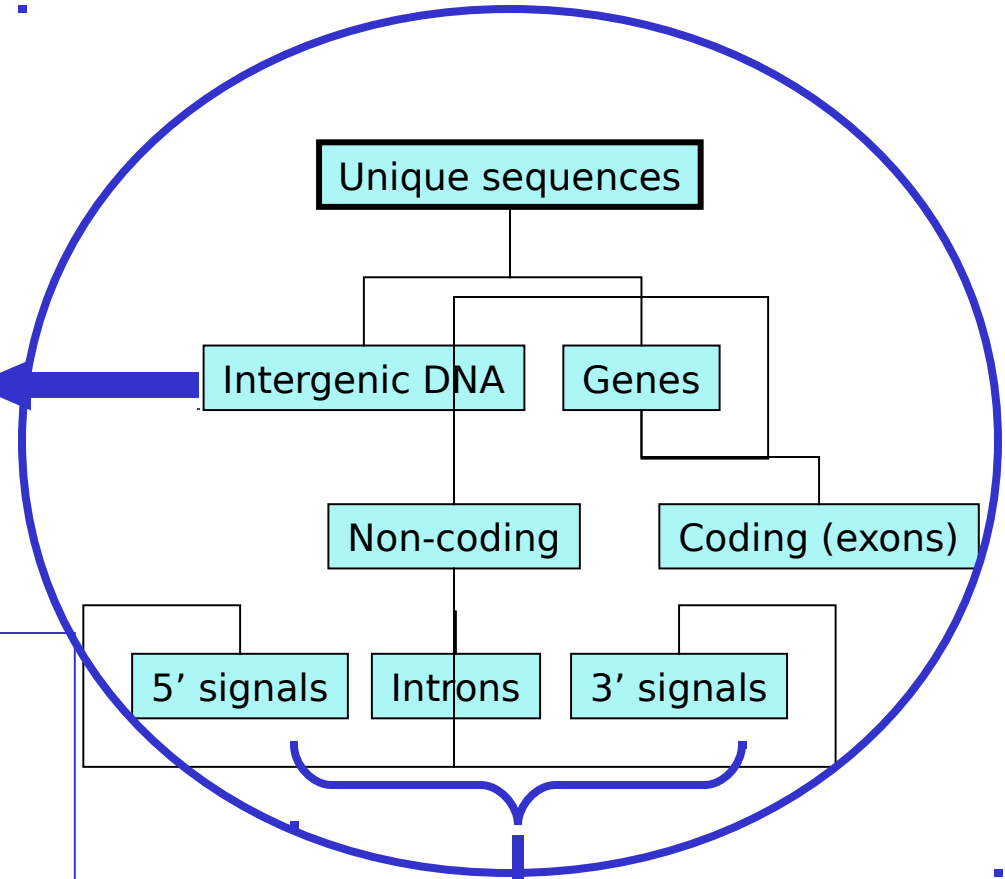
Polymorphism can occur anywhere, but when we consider phenotypic changes, we are normally interested in gene products and gene expression

Includes promoters, enhancers, cis and trans activation domains, regulatory RNA



Control of 'when, where, how much, which type'

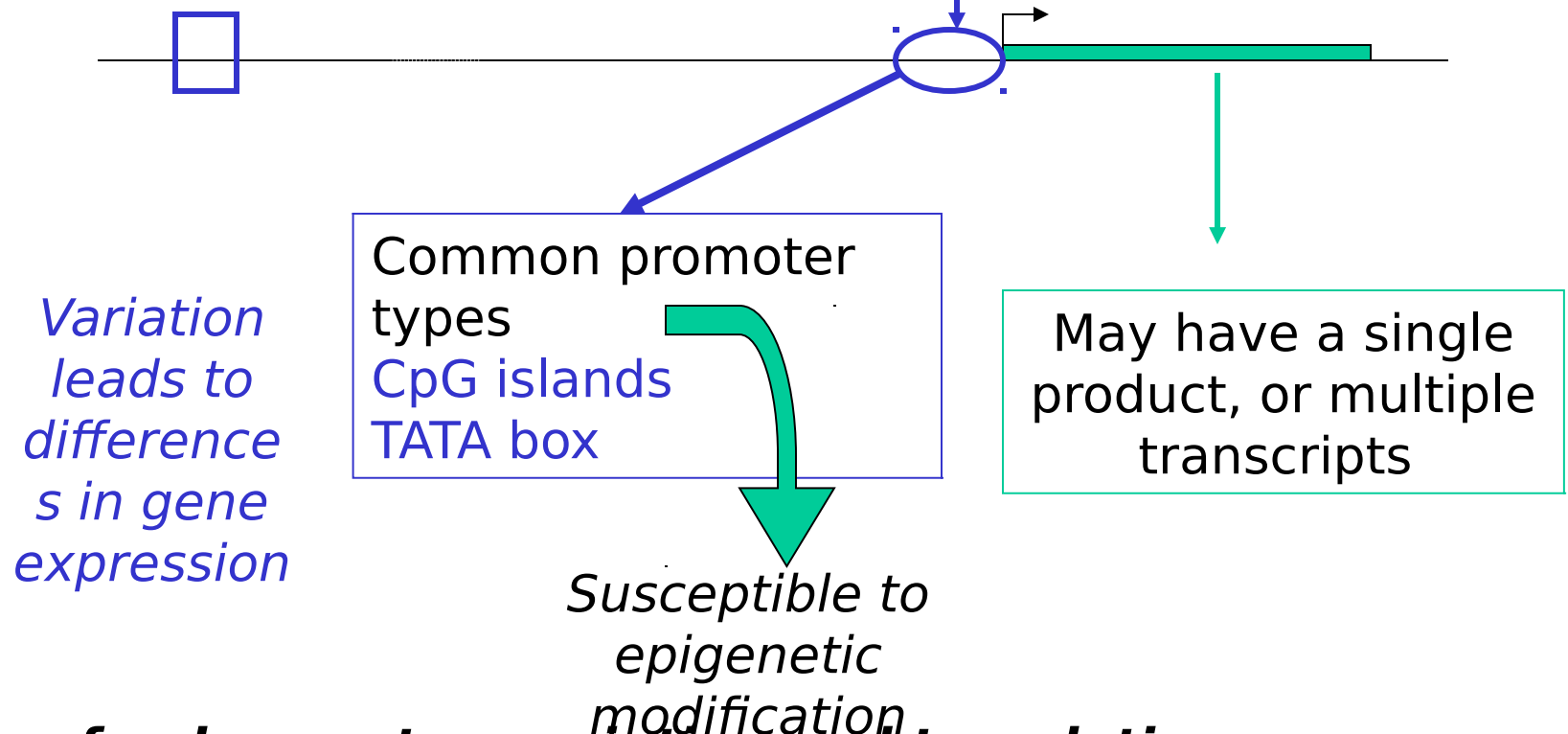
Only a small amount of the genome is protein coding. Most 'functional' DNA will be concerned with the regulation of gene products, and some of the gene products will include non-coding RNAs that in turn refine gene expression



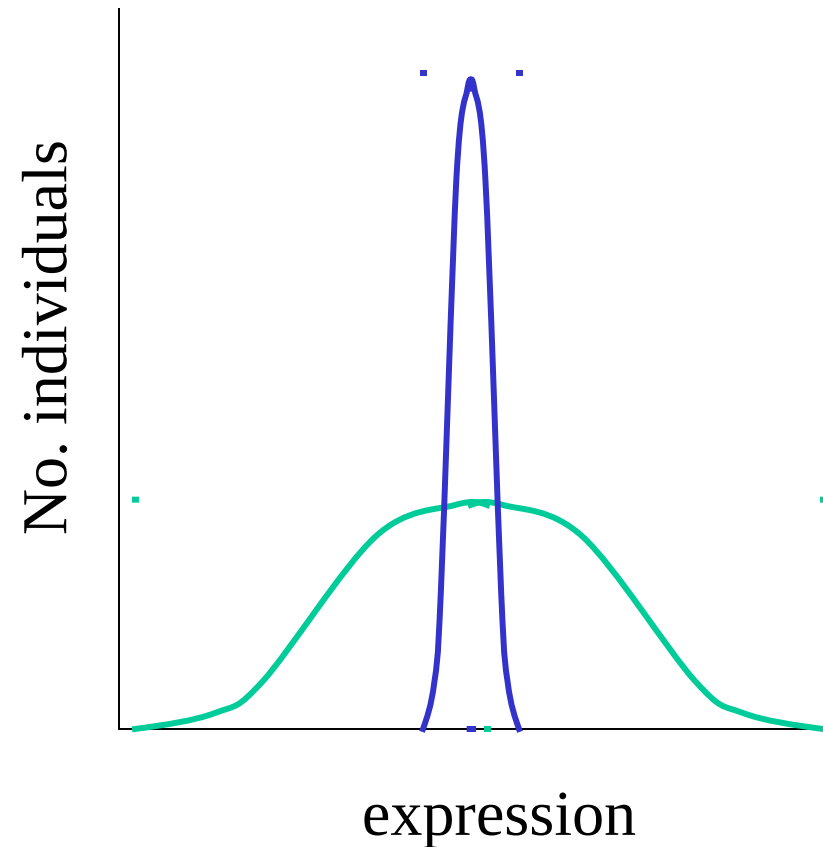
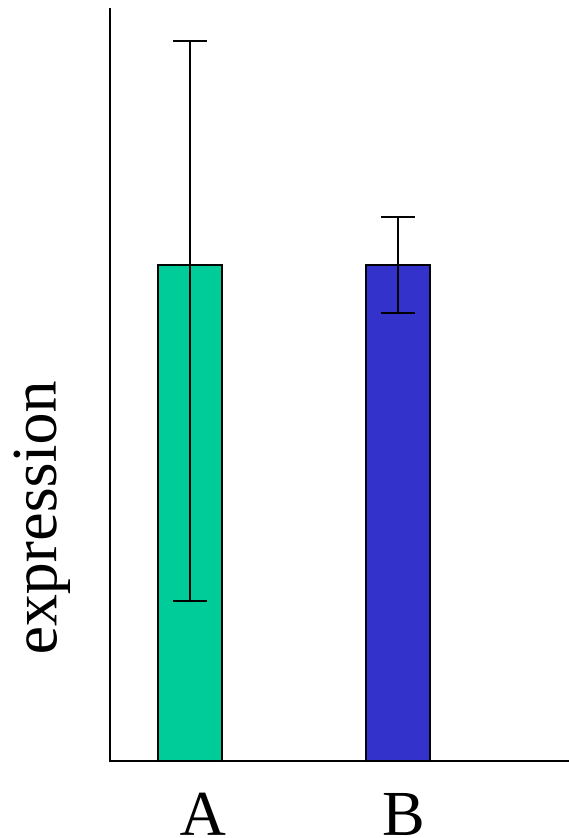
Includes coding potential, splice signals, mRNA stability signals, motifs for post-translational modification

Activators, repressors and enhancers also influence transcription. These can be *cis* (affecting genes nearby) or *trans* (affecting genes elsewhere in genome, maybe different chromosome)

Promoter: immediately upstream of start site of transcription: binds RNA polIII in presence of other transcription factors



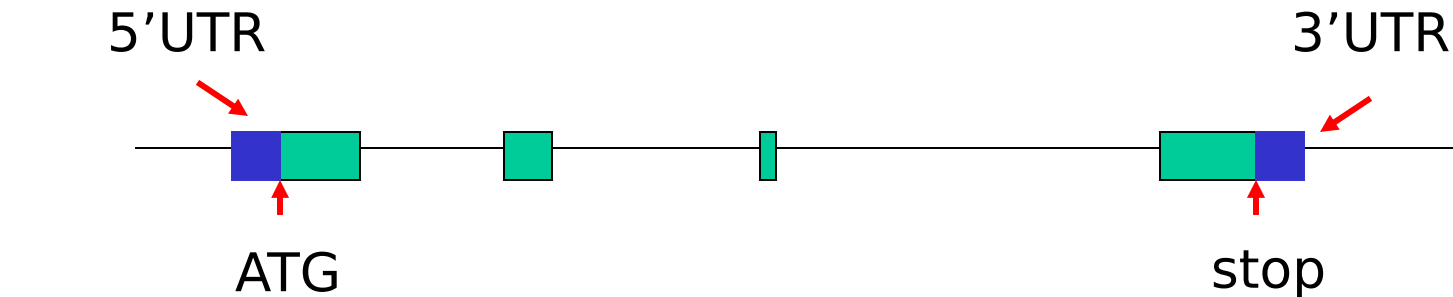
For a refresher on transcription and translation, see any good molecular biology text book



Within a series of normal tissues from normal individuals expression may show high variance or low variance. Polymorphism in regulation?

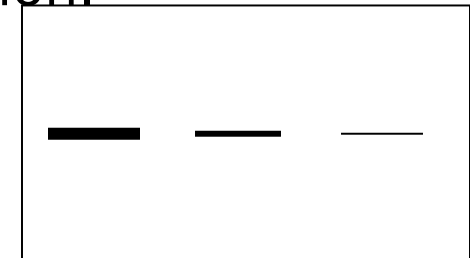
The next few slides summarise material you have met before. Any good textbook will allow you to refresh your memory on transcription and translation.


Changes to expression




← Promoter region →

Motifs in the 3'UTR of transcripts may help to stabilise the mRNA and protect it from degradation.

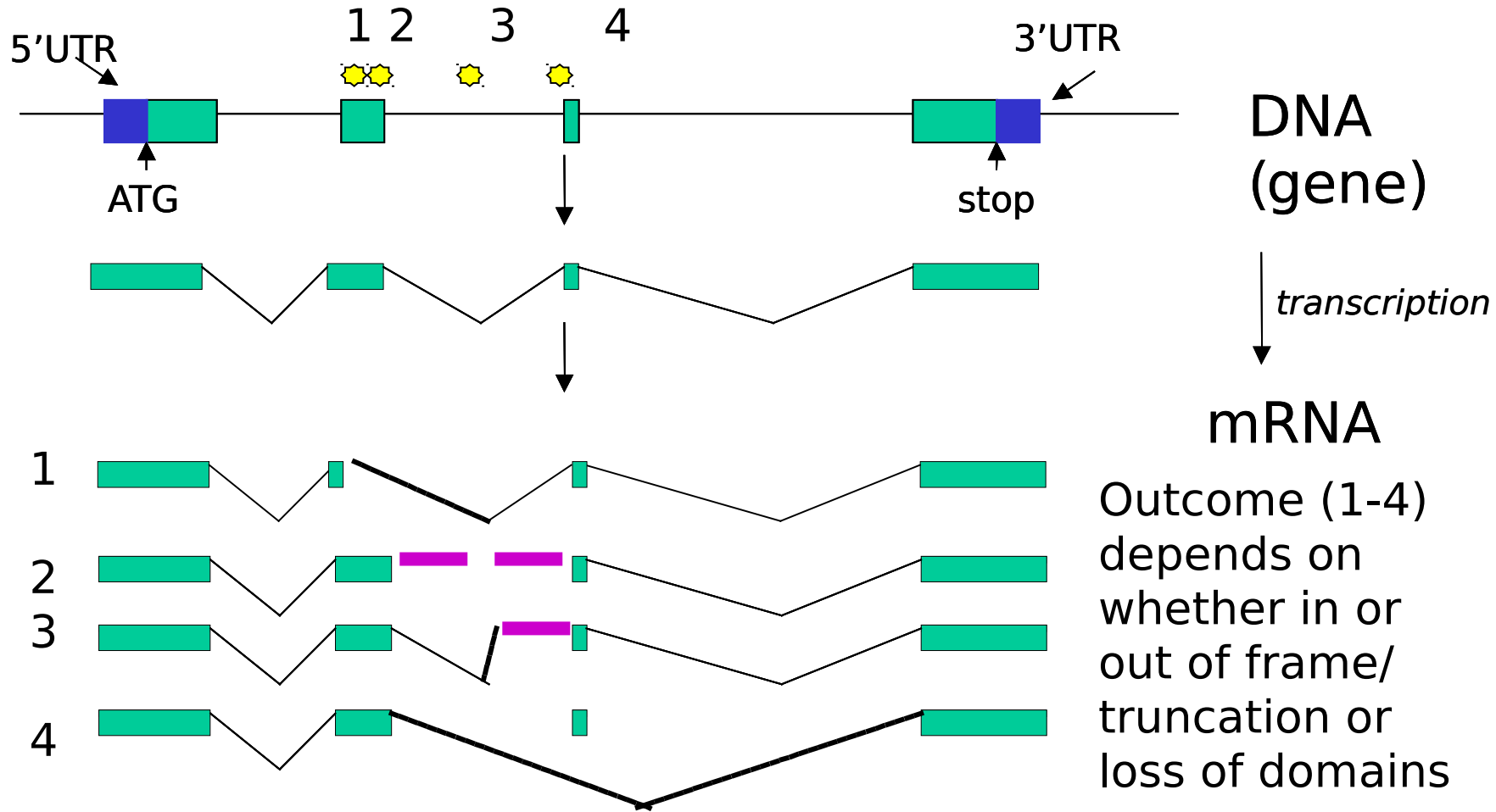



Coding segment of exon


non-coding segment of exon

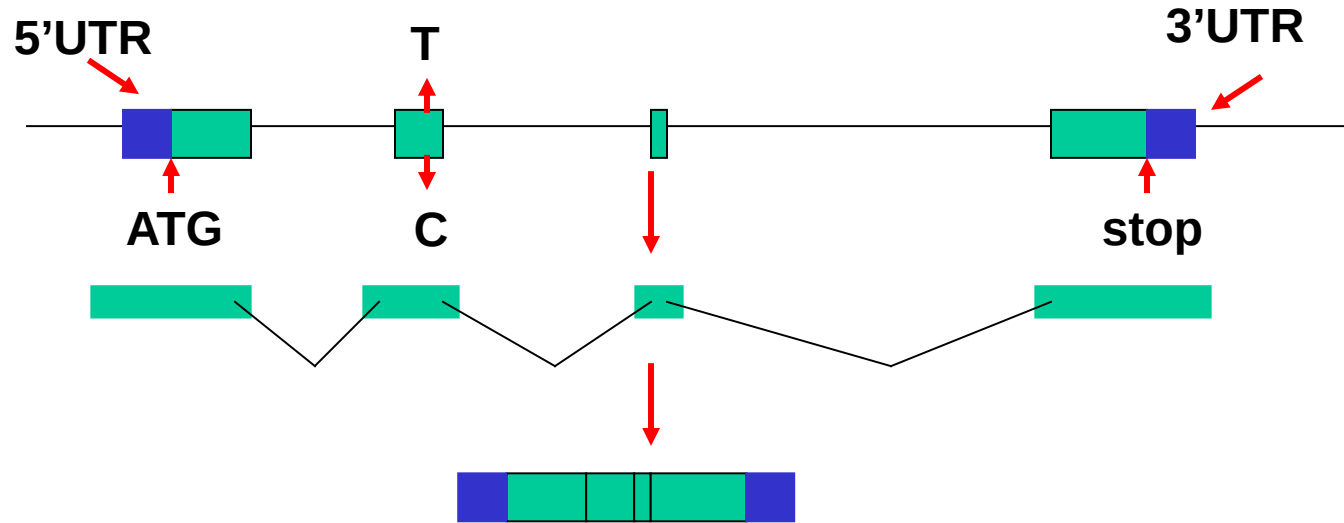
Measure impact of a sequence change at level of mRNA (use northern blot) or protein (use western blot)

Splice site mutations



1. Creation of splice junction in exon
2. Loss of splice consensus at donor site: may insert intronic sequence
3. Creation of new splice consensus site in intron (or loss of internal splice site motifs for correct splicing): may insert intronic sequence
4. Loss of splice consensus at acceptor site: exon skipping

Codon mutation summary



Codon change, but same amino acid: **synonymous**

Codon change, different amino acid: **non-synonymous (missense)**

Non-synonymous changes can be similar (**conservative**) or different amino acid type (**non-conservative**)

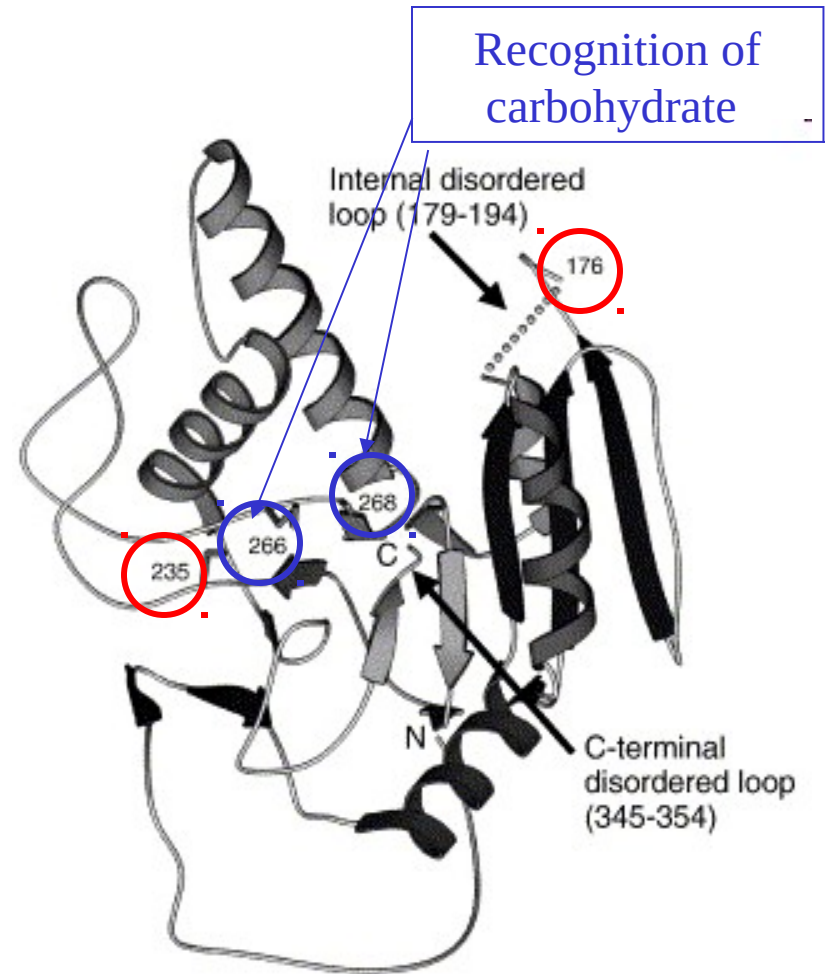
Introduction of stop codon through codon mutation (including small deletions and insertions): **nonsense**

variant	location	effect	Frequency in genome	Risk of phenotypic change
nonsense	Coding	Premature termination	Very low	Very high
Missense (non-synonymous) non-conservative change	Coding	Different amino acid in peptide chain, but similar properties	low	Moderate to high depending on location of amino acid change, e.g. affects binding domain or tertiary structure
Missense (non-synonymous) conservative change	Coding	Different amino acid in peptide chain but similar properties	low	Low to high depending on location of amino acid change
Insertion or deletion with frameshifts	Coding	Changes the frame of the protein coding region, and often has a negative effect	low	Very high depending upon the location of the variant, e.g. results in early termination
Insertion or deletion, in frame	Coding or non-coding	Changes amino acid sequence	low	Low to very high depending upon the amino acids altered
Sense (synonymous)	coding	Does not alter amino acid	medium	Low to high (if splicing is altered)
Promoter/ regulatory region	Promoter, 5'UTR, 3'UTR	Can alter the level, timing or location of gene expression	Low to medium	Low to high depending on the total impact on expression levels
Splice site/ intron exon boundary	Usually within of 10bp boundary	May change splicing pattern or efficiency of introns	low	Low to high depending upon tissue specific patterns of expression
intronic	Deep within intronic sequence	May alter splicing/ expression/ mRNA stability	medium	Very low
intergenic	Non-coding, between genes	Might alter expression through enhancer or other mechanisms	high	Very low

The above is adapted from a review article (Tabor et al. Nature Genetic Reviews 2002) and summarises the hierarchy of risk that a sequence change is a (phenotypic) mutation rather than a polymorphism.

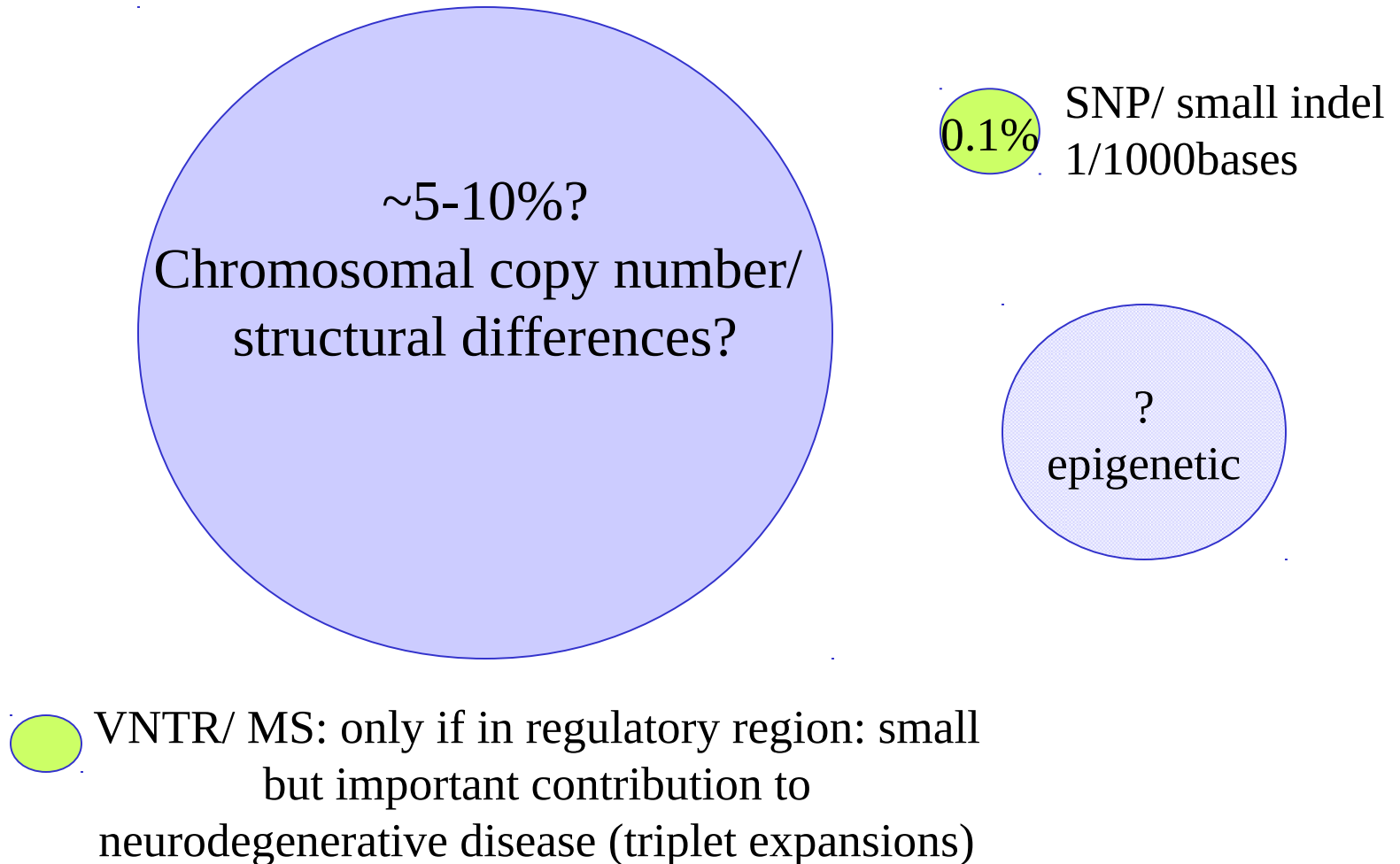
Polymorphism in the coding potential: selection

- ABO blood group
 - Enzyme, glycosyltransferase
 - A transfers N-acetylgalactosamine
 - B transfers galactose
 - Differences at 4 amino acid positions
 - O has no activity, and gene has frameshifts
- AB protective against cholera
- O protective against malaria, lower risk of pancreatic cancer
- *Note that early genetic studies often used phenotypic variation with protein polymorphisms as markers for linkage studies*

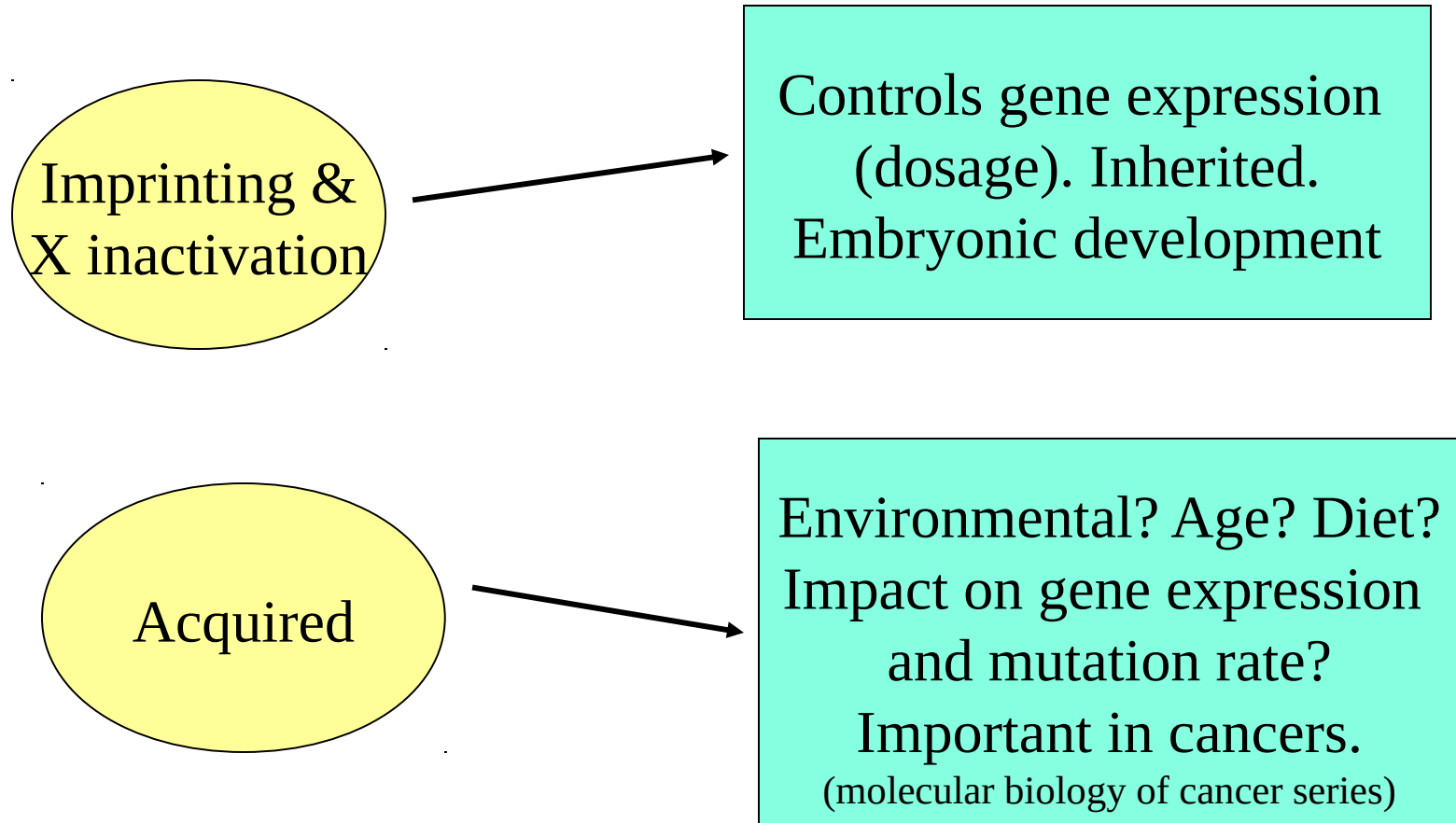


Modified from Yazer and Palcic, *Transfus Med*
Rev. 2005

Contribution to normal phenotypic variation? How much of the genome is involved?



Chromatin level: Epigenetic modification



Most epigenetic models considered in this course will be DNA methylation based, but epigenetic modification may also include histone modifications (>200)

Epigenetic modification

- No change to base sequence
- Modification of histones – may alter chromatin structure and affect gene transcription
 - Detect using antibodies that recognise modified histone proteins (acetylated/ methylated)
- Modification of DNA (often methylation at CpG) again affecting gene transcription.
 - Can detect using specialist PCR; bisulphite modification of DNA followed by either sequencing (specific loci) or array based assay detection (whole genome methylation) to determine ratio of methylated to unmethylated DNA

Covered in more detail later in the course